

A COMPREHENSIVE GEOSPATIAL KNOWLEDGE DISCOVERY FRAMEWORK
FOR SPATIAL ASSOCIATION RULE MINING

by

Thi Hong Diep Dao

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Geography and Urban Regional Analysis

Charlotte

2013

Approved by:

Dr. Jean-Claude Thill

Dr. Wenwu Tang

Dr. Gang Chen

Dr. Robert Brame

ABSTRACT

THI HONG DIEP DAO. A Comprehensive geospatial knowledge discovery framework for spatial association rule mining. (Under direction of DR. JEAN-CLAUDE THILL)

Continuous advances in modern data collection techniques help spatial scientists gain access to massive and high-resolution spatial and spatio-temporal data. Thus there is an urgent need to develop effective and efficient methods seeking to find unknown and useful information embedded in big-data datasets of unprecedentedly large size (e.g., millions of observations), high dimensionality (e.g., hundreds of variables), and complexity (e.g., heterogeneous data sources, space-time dynamics, multivariate connections, explicit and implicit spatial relations and interactions). Responding to this line of development, this research focuses on the utilization of the association rule (AR) mining technique for a geospatial knowledge discovery process.

Prior attempts have sidestepped the complexity of the spatial dependence structure embedded in the studied phenomenon. Thus, adopting association rule mining in spatial analysis is rather problematic. Interestingly, a very similar predicament afflicts spatial regression analysis with a spatial weight matrix that would be assigned a priori, without validation on the specific domain of application. Besides, a dependable geospatial knowledge discovery process necessitates algorithms supporting automatic and robust but accurate procedures for the evaluation of mined results. Surprisingly, this has received little attention in the context of spatial association rule mining.

To remedy the existing deficiencies mentioned above, the foremost goal for this research is to construct a comprehensive geospatial knowledge discovery framework using spatial association rule mining for the detection of spatial patterns embedded in

geospatial databases and to demonstrate its application within the domain of crime analysis. It is the first attempt at delivering a complete geo-spatial knowledge discovery framework using spatial association rule mining.

DEDICATION

To my family

ACKNOWLEDGMENTS

I would like to express my sincere appreciation to my advisor, Dr. Jean-Claude Thill, for a wonderful guidance along this Ph.D. process over the last five and a half years. Not only that I thank you for the inspiration in research that you have created but also for the dedication to student mentoring that you have provided.

I would also like to thank the other members of my dissertation committee: Dr. Wenwu Tang, Dr. Gang Chen, and Dr. Robert Brame for your encouraging support over time and valuable feedbacks on this work.

I deeply acknowledge many of my friends and colleagues in the UNC-Charlotte Geography graduate program for your constructive critiques on this work particularly as well as your delightful friendships over the years.

Last but not least, I am most grateful for the unending support of my family in many different ways. Moms, Dad, and sisters: thank you for your endless love and the enriching guidance through my life. My loving husband, Phong: thank you for your love, your patience, your great support and encouragement which make this dissertation accomplished.

TABLE OF CONTENTS

LIST OF TABLES	x
LIST OF FIGURES	xii
CHAPTER 1: INTRODUCTION	1
1.1. Statement of Research	5
1.2. Significance of the Study	9
1.3. Dissertation Structure	9
CHAPTER 2: FUNDAMENTALS OF SAR MINING AND DISCOVERY	11
2.1 Characteristics and Process of Geographic Knowledge Discovery	11
2.2 AR Mining: Concepts, Algorithms, and Recent Advances	19
2.3 SAR Mining: Challenges and Achievements	26
2.3.1 Definition of SAR	26
2.3.2 Existing SAR Mining and Discovery Algorithms and Frameworks	27
2.3.3 Spatial Predication – The First Remaining Challenge	34
2.3.4 Evaluation and Visualization – The Second Remaining Challenge	41
CHAPTER 3: FUNDAMENTALS OF SPATIAL CRIME ANALYSIS	57
3.1 Techniques for Spatial Crime Analysis	57
3.2 SAR Mining and Discovery: An Alternative	63
3.3 State of Knowledge in Crime Associations	65
3.3.1 Neighborhood Characteristics and Crime	65
3.3.2 Routine Activity and Crime	68
3.3.3 Environment and Crime	70
3.3.4 Thefts of, and from, Motor Vehicles as a Particular Case	74

CHAPTER 4: RESEARCH QUESTIONS	77
4.1 A General Framework for SAR Discovery	77
4.2 Framework Validation	80
CHAPTER 5: SPATIALARMED FRAMEWORK DEVELOPMENT	82
5.1 The SpatialARMED Framework	82
5.2 Mining Spatial Dependence Structure and Modeling Spillover Effects	89
5.2.1 AMOEBA for Spatial Dependence Structure Quantification	90
5.2.2 Spatial Dependence Structure Spillover Effect Model	96
5.3 The Process of Predication	100
5.3.1 Unit of Tuples and Effect of MAUP in SAR Mining	101
5.3.2 Spatial Join Operation	101
5.3.3 Numeric-to-Nominal Mapping Mechanism	103
5.4 Visual Analytics with Subjective Evaluation	105
CHAPTER 6: SPATIALARMED FOR CRIMINOLOGY	113
6.1 Case Study – Dangerous Streets of High Criminal Activities	113
6.2 SpatialARMED Implementation Aspects	123
6.3 SpatialARMED Level 2: Spatial Data Analysis	128
6.3.1 Identifying Spatial Dependence Structures	128
6.3.2 Modeling Spillover Effects of Spatial Dependence Structures	147
6.4 SpatialARMED Level 3: Predication	161
6.4.1 Define Unit of Tuples and Generate Numeric Value Predicates	161
6.4.2 Numeric-to-Nominal Mapping with Crisp Boundary	165
6.4.3 Numeric-to-Nominal Mapping with Fuzzy Boundary	170

6.4.4	Ready-to-mine input files	174
6.5	SpatialARMED Level 4: Mining Rules	177
6.6	Discovered Associations and SpatialARMED Valuation	180
6.6.1	SpatialARMED in Discovering Confirmative Rules	185
6.6.2	SpatialARMED in Discovering Potentially New Rules	197
CHAPTER 7: CONCLUSIONS		220
REFERENCES		224
APPENDIX A:	REPRESENTATIVE MINED SARS FOR DANGEROUS STREETS DUE TO CRIME OF ALL TYPES USING CRISP MAPPING, SUPPORT THRESHOLD OF 5%, CONFIDENCE THRESHOLD OF 40%	245
APPENDIX B:	REPRESENTATIVE MINED SARS FOR DANGEROUS STREET DUE TO MOTOR VEHICLE THEFT USING CRISP MAPPING, SUPPORT THRESHOLD OF 5%, CONFIDENCE THRESHOLD OF 40%	253
APPENDIX C:	REPRESENTATIVE MINED SARS FOR DANGEROUS STREET DUE TO THEFTS FROM MOTOR VEHICLE USING CRISP MAPPING, SUPPORT THRESHOLD OF 5%, CONFIDENCE THRESHOLD OF 40%	256
APPENDIX D:	REPRESENTATIVE MINED FUZZY SARS FOR DANGEROUS STREET DUE TO CRIME OF ALL TYPES USING SUPPORT THRESHOLD OF 5%, CONFIDENCE THRESHOLD OF 40%	263
APPENDIX E:	REPRESENTATIVE MINED FUZZY SARS FOR DANGEROUS STREET DUE TO MOTOR VEHICLE THEFT USING SUPPORT THRESHOLD OF 5%, CONFIDENCE THRESHOLD OF 40%	272
APPENDIX F:	REPRESENTATIVE MINED FUZZY SARS FOR DANGEROUS STREET DUE TO THEFT FROM MOTOR VEHICLE USING SUPPORT THRESHOLD OF 5%, CONFIDENCE THRESHOLD OF 40%	273

LIST OF TABLES

TABLE 1:	A single relational table derived from dataset ABC	89
TABLE 2:	An example of spatial join within the SpatialARMED framework	103
TABLE 3:	Criminal Incident Count in Categories by CMPD National Incident Based Reporting System	116
TABLE 4:	Total incident count by crime of all types and motor vehicle related types by National Incident Based Reporting System (using the Highest Class)	118
TABLE 5:	Data sets	122
TABLE 6:	Identified variables for spatial dependence structure quantification	123
TABLE 7:	Summary of clustering algorithms for AMOEBA and GeoDa	146
TABLE 8:	Process of generating predicates with block group as unit of tuples for the dangerous street mining task (SS stands for Spatial Dependence Structure)	162
TABLE 9:	Summary of predicate types for dangerous street SAR mining	165
TABLE 10:	Summary of numeric-to-nominal mapping mechanism using crisp boundary for all predicates	166
TABLE 11:	Non-uniform membership functions applied to fuzzy mapping of Type 2 predicates	172
TABLE 12:	Predicate name - ID schema table for crisp SAR in the ready-to-mine format	176
TABLE 13:	Statistics on generated frequent item sets and rules for dangerous streets SAR mining	178
TABLE 14:	Library of known associations to crime	181
TABLE 15:	List of confirmative rules for CAT using crisp boundary mapping	188

TABLE 16:	Confirmative rules for CAT using fuzzy boundary mapping	190
TABLE 17:	Listed of confirmative rules for MVT using crisp boundary mapping	192
TABLE 18:	Listed of confirmative rules for MVT using fuzzy boundary mapping	193
TABLE 19:	Listed of all confirmative rules for thefts from motor vehicle using crisp boundary mapping	194
TABLE 20:	Listed of all confirmative rules for thefts from motor vehicle using fuzzy boundary mapping	194
TABLE 21:	Confirmed associations to crime by SpatialARMED	197
TABLE 22:	SpatialARMED strongest rules for CAT	202
TABLE 23:	SpatialARMED strongest rules for MVT	204
TABLE 24:	SpatialARMED strongest rules for TFM	207
TABLE 25:	List of discovering association predicates to crime with fuzzy mapping	210
TABLE 26:	For areas with high homeownership: SpatialARMED strongest discovered associations	212
TABLE 27:	For areas has high income: strongest discovered SpatialARMED associations	212
TABLE 28:	For areas that have high income: SpatialARMED associations	212
TABLE 29:	For areas under strong influence of high income: SpatialARMED discovered associations	212
TABLE 30:	For areas under strong influence of high income: SpatialARMED discovered associations	212
TABLE 31:	For areas under strong influence of high income: SpatialARMED discovered associations	212

LIST OF FIGURES

FIGURE 1:	Geo-spatial knowledge discovery process using spatial association rule mining and roles of geographic domain experts	16
FIGURE 2:	The pseudo Apriori-like algorithm	21
FIGURE 3:	Software architecture of SPADA (Lisi and Malerba 2002)	30
FIGURE 4:	An example of fuzzy concept hierarchy (adapted from Kacar and Cicekli 2002)	31
FIGURE 5:	An example of fuzzy spatial relationship hierarchy (adapted from Kacar and Cicekli 2002)	32
FIGURE 6:	Table-based visualization of association rules	45
FIGURE 7:	2D-matrix based visualization of association rules	45
FIGURE 8:	Three-dimension matrix visualization of association rules	46
FIGURE 9:	Direct graph visualization of association rules	47
FIGURE 10:	Association Rule Network	48
FIGURE 11:	The TwoKey Plot	48
FIGURE 12:	Double-Decker Plot	49
FIGURE 13:	Parallel coordinate plot for association rule visualization	49
FIGURE 14:	Items visualization	52
FIGURE 15:	Rules Visualization	52
FIGURE 16:	Conjoint Visualization	53
FIGURE 17:	The framework of areal categorized geospatial knowledge discovery (adapted from Lee and Phillips 2008)	64
FIGURE 18:	Multivariate association mining technique with (a) ArcGIS feature layers, (b) attribute overlaid for a processing cube, and (c) layer overlaid (adapted from Estivill-Castro 2001; (Estivill-	65

Castro and Lee 2001)

FIGURE 19:	SpatialARMED - a comprehensive framework for SAR mining and discovery	83
FIGURE 20:	Spatial data set ABC for the SpatialARMED framework demonstration	89
FIGURE 21:	AMOEBA procedure (adapted from Aldstadt and Getis 2006)	93
FIGURE 22:	AMOEBA parallel computing workflow	94
FIGURE 23:	AMOEBA parallel processing phase 1 with regional decomposition	94
FIGURE 24:	An example to demonstrate the AMOEBA-based clustering result for the dataset ABC	96
FIGURE 25:	Distance weighted spillover effect models	97
FIGURE 26:	Modeling the spatial spillover impact using cluster G^* values	98
FIGURE 27:	Modeling the spatial spillover impact using individual G_i^* values	99
FIGURE 28:	Normal distribution	105
FIGURE 29:	Pipeline and Components of Visual Analytic Process in SpatialARMED	108
FIGURE 30:	SpatialARMED domain knowledge integrated rule evaluation process using interactive branching approach	109
FIGURE 31:	(a) Study area and (b) Charlotte street network and crime incidents	115
FIGURE 32:	Incident count on occurrence locations of (a) crime of all types (CAT), (b) vehicle theft (MVT), and (c) theft from vehicle (TFM)	119
FIGURE 33:	Incident count on associated street blocks of (a) crime of all types (CAT), (b) motor vehicle theft (MVT), and (c) theft from motor vehicle (TFM)	120

FIGURE 34:	Representative facilities as potential crime generators and attractors in the City of Charlotte	122
FIGURE 35:	Spatial implementation work flow for case study	127
FIGURE 36:	AMOEBA-based hot spots for crime of all types (CAT)	130
FIGURE 37:	AMOEBA-based hot spots for motor vehicle thefts (MVT)	131
FIGURE 38:	AMOEBA-based hot spots for thefts from motor vehicle (TFM)	132
FIGURE 39:	AMOEBA-based clusters for per capital income using ACS2010 BLG data with Rook neighborhood definition	133
FIGURE 40:	AMOEBA-based clusters for percentage of population with high school degree or above using ACS2010 BLG data with Rook neighborhood definition	134
FIGURE 41:	AMOEBA-based clusters for percentage of multiple (> 2) unit homes using ACS2010 BLG data with Rook neighborhood definition	135
FIGURE 42:	AMOEBA-based cluster for percent of population who rent and move in less than 5 years using ACS2010 BLG data with Rook neighborhood definition	136
FIGURE 43:	AMOEBA-based clusters for percentage of employed population using ACS2010 BLG data with Rook neighborhood definition	137
FIGURE 44:	AMOEBA-based clusters for ethnic heterogeneity index using Census 2010 BLK data with Rook neighborhood definition	138
FIGURE 45:	AMOEBA-based clusters for percentage of African-American population using Census 2010 BLK data with Rook neighborhood definition	139
FIGURE 46:	AMOEBA-based cluster for percentage of owner occupied houses using Census 2010 BLK data with Rook neighborhood definition	140
FIGURE 47:	AMOEBA-based clusters for percentage of males aged 18-24 using Census 2010 BLK data with Rook neighborhood	141

	definition	
FIGURE 48:	AMOEBA-based clusters for percentage of single-parent families using Census 2010 BLK data with Rook neighborhood definition	142
FIGURE 49:	AMOEBA-based hotspots for business activities using 2008 business data	143
FIGURE 50:	Percentage of AA population: (a) data statistics; (b) AMOEBA clustering result; (c, d) GeoDa Gi* clustering results using normal p value and pseudo p value respectively	145
FIGURE 51:	AMOEBA-based detected streets of high crime (left) and spatial spillover effect (right)	148
FIGURE 52:	AMOEBA-based detected streets of high motor vehicle thefts (left) and spatial spillover effect (right)	149
FIGURE 53:	AMOEBA-based detected streets of high thefts from motor vehicle (left) and spatial spillover effect (right)	149
FIGURE 54:	AMOEBA-based detected streets of high commercial activity (left) and spatial spillover effect (right)	150
FIGURE 55:	Per Capital Income: AMOEBA-based detected clusters of high values (upper left) with its spatial spillover effect (upper right) and AMOEBA-based	151
FIGURE 56:	Percentage of population with high school degree or above: AMOEBA-based detected clusters of high values (upper left) with its spatial spillover effect (upper right) and AMOEBA-based	151
FIGURE 57:	Percentage of housing structures of 3 or more units: AMOEBA-based detected clusters of low values (bottom left) with its spatial spillover effect (bottom right)	151
FIGURE 58:	Percentage of population who rent and move in less than 5 year: AMOEBA-based detected clusters of high values (upper left) with its spatial spillover effect (upper right) and AMOEBA-based	151

FIGURE 59:	Percentage of population employed: AMOEBA-based detected clusters of high values (upper left) with its spatial spillover effect (upper right) and AMOEBA-based	151
FIGURE 60:	Heterogeneity Index: AMOEBA-based detected clusters of high values (left) and spatial spillover effect (right)	151
FIGURE 61:	Percentage of African-American population: AMOEBA-based detected clusters of high values (left) and spatial spillover effect (right)	151
FIGURE 62:	Percentage of owner occupied homes: AMOEBA-based detected clusters of high values (left) and spatial spillover effect (right)	151
FIGURE 63:	Percentage of population who are males aged 18-24: AMOEBA-based detected clusters of high values (left) and spatial spillover effect (right)	151
FIGURE 64:	AMOEBA-based detected cluster of high percentage of single-parent families (left) and spatial spillover effect (right)	158
FIGURE 65:	Location of hotels and models (left) and spatial spillover effect (right)	158
FIGURE 66:	Location of Wal-Mart super centers (left) and spatial spillover effect (right)	159
FIGURE 67:	Location of shopping malls (left) and spatial spillover effect (right)	159
FIGURE 68:	Location of park-and-ride facilities (left) and spatial spillover effects (right)	160
FIGURE 69:	Location of Alcohol Drinking Places (left) and spatial spillover effects (right)	160
FIGURE 70:	Standardized predicate values histogram and crisp boundary mapping mechanism for Type 1 predicates (Block group variables)	167
FIGURE 71:	Standardized predicate value histogram and crisp boundary mapping mechanism for Type 2 predicates (Block and Street	168

variables)	
FIGURE 72: Crisp boundary mapping mechanism for Type 3 predicates (SIM based values) using standardized values	169
FIGURE 73: Format of final relational table to mine using crisp mapping	169
FIGURE 74: Standardized value histogram and fuzzy boundary mapping mechanism for Type 2 predicates	171
FIGURE 75: Fuzzy boundary mapping mechanism for Type 3 predicates using standardized values	174
FIGURE 76: A record in the final ready-to-mine relational table with crisp mapping	175
FIGURE 77: A record in the final ready-to-mine relational table with fuzzy mapping	175
FIGURE 78: Process of mining SAR with LUCS KDD ARM	179
FIGURE 79: Visualization of confirmative SAR for CAT with (a) crisp and (b) fuzzy boundary mapping	187
FIGURE 80: Visualization of confirmative SAR for MVT with (a) crisp and (b) fuzzy boundary mapping	191
FIGURE 81: Visualization of all confirmative SAR for TFM with (a) crisp and (b) fuzzy boundary mapping	196
FIGURE 82: The strongest rules for (a) high and (b) medium high CAT	203
FIGURE 83: The strongest rules for high MVT	205
FIGURE 84: The strongest rules for high (a) and medium high (b) TFM	208
FIGURE 85: Potentially new fuzzy rules for TFM in association with high homeownership	212
FIGURE 86: Fuzzy rules for TFM for areas has high income and medium spillover impact of TFM	214
FIGURE 87: The strongest fuzzy rules for crime of all type for areas under	217

strong spillover effect of high income

FIGURE 88: Fuzzy rules for TFM for areas under strong spillover impact
high income

218

CHAPTER 1: INTRODUCTION

Geographic information science has been experiencing an unprecedented increase in data accessibility and computational capability (Miller 2007; Mennis and Guo 2009; Miller and Han 2009). More than ever before, continuous advances in modern data collection techniques help spatial scientists gain access to massive and high-resolution spatial and spatio-temporal data. High spatial, temporal, and spectral resolution remote sensing systems offer substantial volumes of geo-referenced digital imagery and video. Location-aware technologies based on freely available Global Positioning System (GPS) and Galileo satellites or cellular and Wi-Fi signals that have emerged into applications of wireless cell-phones, in-vehicle navigation systems and wireless internet clients allow researchers to collect non-stop unprecedented amounts of data for both outdoor and indoor environments at the level of individual movement. Moreover, information infrastructure initiatives such as the U. S. National Spatial Data Infrastructure are easing data sharing and interoperability (Miller 2007). Geographic information scientists are not unaware of this trend. The 2012 GIScience conference organized in Columbus, OH had a dedicated theme on “Big data” along with a series of workshop presentations and panel discussions on the definition of big (geospatial) data itself, as well as the research opportunities and challenges it brings. The term ‘big data’ is actually not new. Laney (2001) initially associated it with three V’s characteristics: Volume (i.e. high in volume, in the range of petabyte sizes), Variety (i.e. heterogeneous in data types), and Velocity

(i.e. high velocity, near real time or real time, data acquisition). The definition of big data sometimes is extended with the addition of new Vs, such as V for Virtual referring to online data assets, V for veracity, and V for value, and V for visualization. As the data volume, variety, and velocity increase, its veracity and value also increase. Particularly for geographic information scientists, the big data era promises major new opportunities to gain a better understanding of important and complex geographic phenomena, such as human-environment interaction, socioeconomic dynamics, or the interaction between physical and virtual worlds. The challenges, however, involve more than just managing data in high volume and of heterogeneous nature. In order to achieve innovation and productivity, “new forms of processing to enable enhanced decision making, insight discovery and process optimization” are called for (Laney 2012).

Traditional spatial statistics methods were designed in a data-poor era with limited computing power (Miller and Han 2009). They are thus most efficient to handle small, scientifically sampled and homogenous datasets. When dealing with a significant increase in data volume and diversity in the nature of digital geographic data, they reveal some limitations. Firstly, most methods adhere to a limited perspective, such as univariate spatial autocorrelation, or a specific and simple type of relation models, such as linear regression. They are also confirmatory techniques and require a priori hypotheses. Secondly, many of them cannot process a very large data volume. Computing requirements and their confirmatory nature indeed prevent these approaches from discovering unexpected or surprising information (Miller and Han 2001). Thirdly, newly emerging data types, ranging from trajectories of moving objects, to geographic information embedded in webpages and surveillance videos, along with new applications

demand more efficient and effective approaches to discover interesting patterns and information (Mennis and Guo 2009).

As a result, there is an urgent need to develop effective and efficient methods seeking to find unknown and useful information embedded in the big-data datasets of unprecedentedly large size (e.g., millions of observations), high dimensionality (e.g., hundreds of variables), and complexity (e.g., heterogeneous data sources, space–time dynamics, multivariate connections, explicit and implicit spatial relations and interactions) (Mennis and Guo 2009). This has broadly motivated the enlargement of the traditional data mining and knowledge discovery approaches for spatial applications, and consequently leads to the emergence of the so-called spatial data mining and geographic knowledge discovery (GKD) field.

Responding to this line of development, this research focuses on the utilization of the association rule (AR) mining technique for a geospatial knowledge discovery process. The association rule mining approach (Agrawal et al. 1993; Agrawal and Srikant 1994) has been extensively applied in market basket transaction analysis with traditional relational databases aiming to find associations among customer checked-out items, for marketing purposes. It is regarded as one of the most popular mining techniques used for knowledge discovery because it is very efficient in finding frequent and meaningful relations, positive associations and stochastic plus asymmetric patterns in large relational data warehouse. Association rules are therefore promising for spatial data analysis in order to discover unknown spatial patterns, especially for large spatial databases.

Unique characteristics of spatial data, such as spatial attributes, spatial relations, spatial correlations, and spatial hierarchy, however, prevent a direct deployment of an

association rule mining approach in geospatial analysis problems. Association rule mining has been adapted to spatial analysis, and often named spatial association rule (SAR), by simply including spatial predicates (Koperski and Han 1995). With linguistic expressions, spatial predicates allow flexibility in representing explicit spatial relations of objects in terms of distance, direction, and topology but also implicit spatial dependencies, i.e. spatial autocorrelation, or more generally the spatial dependence structure embedded in the studied phenomenon. However, the dynamics and complexity of the spatial component captured by spatial predicates are often overlooked. Moreover, there exists no comprehensive procedure for generating predicates to capture the spatial dependencies. Short of this, adopting association rule mining in spatial analysis is rather problematic. Interestingly, a very similar predicament afflicts spatial regression analysis with a spatial weight matrix that would be assigned a priori, without validation on the specific domain of application.

Besides, similar to any other domain-specific data mining framework, a dependable geospatial knowledge discovery process necessitates algorithms supporting automatic and robust but accurate procedures for the evaluation of mined results. Surprisingly, this has received little attention in the context of spatial association rule mining. This challenge is particularly relevant because the association rule mining approach is well-known for producing a large number of rules, which makes assessment difficult and point to the importance of visual analytics. The existing literature identifies remaining challenges in establishing subjective criteria based on geospatial domain knowledge for evaluating the interestingness of spatial association rules. In addition, possibilities for integrating

multidimensional and interactive geovisualization tools with these criteria for visual analytic purposes have not yet been examined.

1.1. Statement of Research

To remedy the existing deficiencies mentioned above, the foremost goal of this research is to construct a comprehensive geospatial knowledge discovery framework using spatial association rule mining for the detection of spatial patterns embedded in geospatial databases and to demonstrate its application within the domain of crime analysis.

A spatial database utilizing an entity-based (sometimes also referred to as object-based or feature-based) data model is taken into consideration for the mining task. Basically, it represents real-world features, such as countries, counties, census block groups, cities, land parcels, houses, schools, crime incidents, police stations, etc., using the vector-based spatial representation comprised of points, lines, and polygons. A set of attributes (or variables) is associated with each of these features. Attributes can be categorized into *aspatial* (also referred to as semantic) and *spatial* attributes. Semantic attributes are related to non-spatial information of the features such as name, age, price, etc. while spatial attributes are related to spatial characteristics of the feature itself (e.g. size and shape) or the spatial relationships (e.g. being close-to or far-away) to other features. Homogeneous collections of features having the same spatial representation and a common set of attributes are grouped into feature classes. The proposed mining and discovery framework aims at the detection of associations between the *reference or main feature class* and some *task-relevant or associative feature classes*. The reference feature class is the main subject of the rule description and often resides on the right-hand side of

the rules (consequents), while task-relevant ones are both aspatially and spatially relevant for the task at hand and often reside on the left-hand side of the rules (antecedents). For instance, one may be interested in finding associations among expensive houses and other geospatial objects such as mountains, beaches, road network, etc. In this case, the reference feature class or object class is a point feature class representing the locations of expensive houses while the task-relevant ones are polygon and polyline feature classes representing mountains, beaches, and roads in the study area. Following a single relational database mining approach, information is collapsed into a single table whose rows are equivalent to units of mining (tuples) and columns are to semantic and spatial attributes. The mining task, in this case, focuses on finding associations between the *reference or main attribute* and *task-relevant or associative attributes*. For example, for the previously described mining task, a single table containing records for all houses in the study area with attributes being the price of the house, distances to mountains, to beaches, to closest main roads are created. The house price in this case is the reference attribute while others are task-relevant attributes. An example of the expected association rules in this case can be “If a house is very close to the beach then it is expensive”.

Fundamentally, a comprehensive framework for spatial association rule mining and discovery can be decomposed into the following tasks:

- a) Identify associative features
- b) Select and transform semantic attribute information; derive non-spatial predicates
- c) Identify and quantify spatial components involved; derive spatial predicates
- d) Mine spatial association rules

- e) Visualize and evaluate intermediate mined results for interestingness using geospatial knowledge base; update geospatial knowledge base.

As discussed above, the most significant remaining challenges lie with tasks (c) and (e) while others are rather straightforward or well documented already.

Being prefixed by the term “spatial”, this framework should be centered on its capability to handle the spatial component of the problem at hand, namely heterogeneity and dependence of the phenomenon under study across space, as well as spatial interactions among features. Moreover, from a knowledge discovery perspective, it is essential to perform an evaluation. Proficiently addressing these challenges ensure the successful construction of a complete geospatial knowledge discovery framework which can be applied toward spatial problems.

In order to develop the above framework, three major tasks are identified for this research:

The first is to identify and quantify spatial components of the problem at hand. The spatial components referred in this study include not only spatial relationships (i.e. distance-based, topological-based, directional-based, etc.) among studying features but also to spatial dependency as the most important characteristic of spatial processes. Spatial dependency, if it exists, forms a spatial pattern of High or Low value concentrations, which is referred to as *spatial dependency structure* in this study. This is sometimes also referred to as *High/Low clusters*, or *hot/cold spots*. In many cases, SAR mining aims to identify strong associations between High/Low value of one variable and High/Low value of other variables. While most of the existing SAR mining approach apply predetermined concepts of the spatial dependency structures (i.e. subjective or

predetermined concepts of High or Low), this study proposes that the identification of these structures should be driven by the data and this should be performed as an important step within the SAR mining and discovery process. How to quantify and represent spatial dependencies, in fact, remains to be an on-going issue in spatial analysis as because semantics and vagueness are often involved. Particularly, popular but modest means to account for spatial dependence such as global indicators based on simple statistics (e.g. average of differences to the mean) or regular neighborhood structures are undesirable. In this study, the spatial dependency structure for each involved variable are mined to identify their locations, boundaries, sizes, shapes, and concentration magnitude measures. As a result, definitions for the necessary spatial hierarchical concepts or for what is so-called High or Low can be derived objectively, which is fundamental in forming effective spatial association predicates in support of the mining step. After successfully identifying the spatial dependency structure for each variable, another spatial component that should be considered in SAR mining is the proximity effects of these spatial dependence structures on the phenomenon under study. One could relate this to the concept of *spatial spillover impact* of, in this case, the hot/cold spots. Put into the context of SAR mining, considering the spatial spillover impact of hot/cold spots allows looking into indirect spatial functional associations.

The second is to develop a spatial predication mechanism for spatial association rule mining. A procedure that transforms quantitative to linguistic measurements for spatial components is required. Optimal choices should allow fuzzy-set mapping and prioritize automatic procedures.

The third is to construct visual analytic functionality coupled with a geospatial knowledge-based scheme for the subjective evaluation of results of spatial association rule mining. As a large number of rules are typically generated under text format, it would be fastidious to evaluate them without visual analytics. Capability to quickly identify and depict strong and useful associations is the objective of the visualization-evaluation system. The process of detecting potential new and interesting rules should be regulated based on libraries of known associations constructed on domain-specific theories and ontologies.

Once the framework is developed, validation is the subsequent essential task. The proposed framework will be implemented and tested using the case of crime analysis in the city of Charlotte, North Carolina. The proposed approaches seek to identify spatial variations and dependencies of crime across the city, and later to discover interesting associative factors that contribute to its spatial dynamics.

1.2. Significance of the Study

Spatial association rule mining can be effective for extracting unknown patterns within large spatial databases only under the condition that spatial components are well addressed. This study proposes a comprehensive framework and a library of algorithms of spatial analysis and visual analytics to resolve this fundamental challenge and offer practical evaluation tools. The framework is the first attempt in delivering a complete geo-spatial knowledge discovery framework using spatial association rule mining.

1.3. Dissertation Structure

The dissertation is structured as follows. Section 2 provides a literature review on association rule mining and discovery for geospatial databases, focusing on existing

algorithms as well as visual analytic techniques and their adaptation to geospatial data. Section 3 facilitates understanding on literature of crime theories and analysis from spatial perspectives. Next, in Section 4, the specific research questions are stated and discussed. Section 5 discusses the spatial association rule mining and discovery framework. Section 6 presents the practice of SpatialARMED in criminology to mine associations to high crime in Charlotte, NC. Chapter 7 concludes this research, summarizes its contributions, and discusses its limitations and directions for future work.

CHAPTER 2: FUNDAMENTALS OF SAR MINING AND DISCOVERY

This chapter reviews the fundamentals of spatial data mining and the geographic knowledge discovery process along with more recent developments, as well as remaining challenges of association rule-based approaches in both aspatial and spatial contexts.

2.1 Characteristics and Process of Geographic Knowledge Discovery

In the literature, knowledge discovery in databases (KDD) refers to an iterative process involving multiple steps, including data selection, data pre-processing, analysis with computational algorithms (i.e. mining), interpretation and evaluation of the results, formulation and update of pre-existing knowledge bases, adjustment to data and analysis methods, evaluation of result again, and so on (Fayyad et al. 1996). Data mining, on the other hand, is only one step of the knowledge discovery process and is narrowly defined as the application of computational, statistical or visual methods. As the data mining step involves the deployment of techniques to distil data into information implied by the data, the knowledge discovery process entails the higher level process of purifying the mined information into knowledge and beliefs about the world described by the data. At variance with data mining, knowledge discovery requires human intelligence to guide the process and to evaluate the results based on pre-existing knowledge (Miller and Han 2001). Being exploratory and inductive in nature, data mining and knowledge discovery seek to find patterns that are valid (i.e. a generalized pattern, not a data anomaly), novel (i.e. unexpected), useful (i.e. relevant), and understandable (i.e. interpretable and

installable into knowledge) (Fayyad et al. 1996). Generally, the requirement for novelty distinguishes data mining from traditional statistics, which are more oriented towards hypothesis confirmation than generation (Miller 2007).

The universal goal of spatial data mining and geographic knowledge discovery (GKD) is to advance data mining and knowledge discovery methods to analyze large and complex spatial data. The “spatial” term used in this research refers to geospatial data in which data objects are georeferenced; the space concept is embedded in locations on or near the Earth’s surface.

The nature of the geographic space, the complexity of spatial object relationships, the heterogeneous and sometimes ill-structured nature of geographic data, and the individuality of geographic knowledge bring uniqueness into spatial data mining and geographic knowledge discovery, and at the same time, render standard KDD techniques inefficient (Shekhar et al. 2003b). All these points are now discussed in more detail.

First, spatial objects, by definition, are embedded in a continuous space that serves as a measurement framework for all other spatial attributes. This means that they are characterized by a geometric representation and referenced position; the former implicitly defines a number of spatial attributes, while the later defines spatial relations of different nature, such as distance, directional, and topological relationships, not explicitly encoded in a spatial database. Modeling these implicit spatial properties in order to associate them with a clear semantic and a set of efficient procedures for their computation is the first challenge of spatial data mining. This particularly becomes true when there are various human geographic processes that exhibit non-Euclidean spatial properties such as social networking relations, migration behaviours, disease propagation, and cognitive

accessibility, to name a few. Exploring alternative geo-spaces for representing geographic data in these cases is necessary to substantially enhance the GKD process (Miller and Wentz 2003).

Secondly, geographic data often exhibit spatial dependence according to the first law of geography and spatial heterogeneity. This violates the fundamental assumption of classical association rule mining that every items and transactions are independent. Spatial dependence is the tendency of observations that are more proximal in the geographic space to exhibit greater degrees of similarity (i.e. positively auto-correlated) or dissimilarity (negatively auto-correlated). Proximity can be defined in very general terms, involving measures of distance, direction, and/or topology. Spatial heterogeneity is regarded as a non-stationary process with respect to location, i.e. localization. Although spatial dependence and heterogeneity are sometimes caused by misspecification (e.g. missing variables), they reflect the inherent nature of geographical processes and should not be overlooked (Miller and Wentz 2003; Shekhar et al. 2003b).

Thirdly, while multidimensional data objects in typical KDD can be reduced to points without information loss, this is often not the case with GKD. This is because spatial characteristics of objects such as size and shape can have significant influence on the process under study. Moreover, in geographic data, aggregated spatial units, such as census districts, are often used for administrative or confidentiality reasons. This causes a problem in spatial analysis known as the modifiable areal unit problem (MAUP) (Fotheringham and Wong 1991) and should be carefully considered within the GKD process to determine if a discovered pattern is robust or simply an artifact of the spatial measurement units (Miller 2007).

Fourthly, spatio-temporal processes introduce additional complexity to the GKD process. For instance, at variance with space, time is directional, has unique scaling and granularity properties, and can be cyclical and even branching with parallel local time streams (Roddick and Lees 2001). Digital geographic data also include more heterogeneous data types, even ill-structured data. In addition to the traditional data models such as vector and raster, geo-referenced data in the form of dynamic flows and space-time trajectories, or georeferenced multimedia in the form of audio, imagery, video and text are increasingly widespread due to the popularity of real-time environmental monitoring systems such as intelligent transportation systems and location-based services (Miller 2007). Information about places and times contained within these data can be very useful to better understand geospatial reality.

Finally, geographic knowledge base can play an important role in guiding and managing the GKD process. If geographic concept hierarchies are used, spatial objects are organized in hierarchies of classes. This means that by descending or ascending through a hierarchy, it is possible to view the same spatial object at different levels of abstraction (or granularity). This leads to the issue of the confidence with which patterns are more likely to be discovered at low granularity levels, whereas large support is more likely to exist at higher granularity levels. Appropriate data mining techniques thus need to be able to explore the search space at different granularity levels. These distinctive characteristics of spatial data must be carefully considered and well addressed for a successful deployment of data mining and knowledge discovery in spatial data warehouse (Koperski et al. 1996; Miller and Han 2001; Shekhar et al. 2002; Shekhar et al. 2003b; Malerba et al. 2009).

Spatial data mining and geographic knowledge discovery are not a push button task (Mennis and Guo 2009). Although it is data-driven, it is also a human-centered, iterative and inductive learning framework in which the user controls for the selection and integration of data, cleaning and transformation of the data, choice of analysis methods, and finally results interpretation. Solutions to the questions of what variables should be selected, what measurement framework should be used, what spatial relations or contextual information should be considered, and whether the chosen data adequately represent the complexity and nature of the problem, should be carefully deliberated. Geographic domain experts or spatial scientists thus can significantly contribute to enhance the process. Figure 1 provides a complete picture of the geographic knowledge discovery process, in which the roles of spatial scientists are clearly articulated. In this figure, the components featuring a role for geographic experts are identified in three areas and highlighted in red. Their first role is to build a spatial data warehouse given a priori constructed knowledge base. The second role of geographic domain experts is to assist spatial data mining as a central component of the geographic knowledge discovery process; and the third is to improve evaluation schemes using existing domain theories and knowledge.

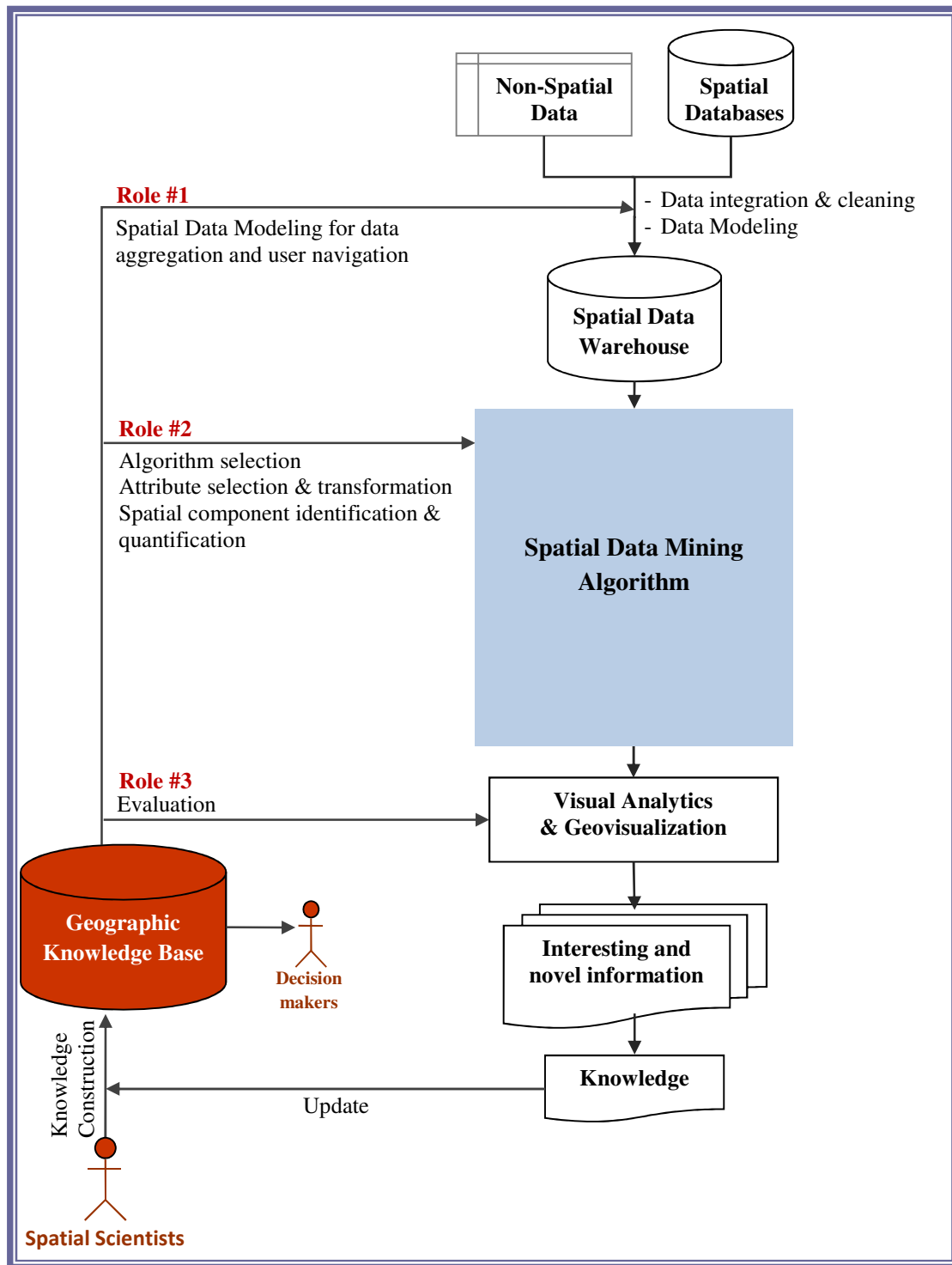


Figure 1: Geo-spatial knowledge discovery process using spatial association rule mining and roles of geographic domain experts

The first component of a GKD process includes spatial data warehouses (SDWs) that contain both spatial (i.e. geo-referenced) and aspatial (i.e. semantic) data. The U.S. Census database is an example of a spatial data warehouse (Shekhar and Chawla 2003). Spatial data warehousing mainly focus on the development of extended and spatial multidimensional data models to support spatial data aggregation and user navigation. Spatial data aggregation is an operation providing a synthetic view of a geographical phenomenon, whereas navigation is a functionality allowing the users to interactively inspect and analyze data through a set of spatially aware operators (Spatial OLAP) (Bertino and Damiani 2005). It is important to note that, while traditional OLAP methods generate summary across tabs in tables with clear standards for aggregation and cross-tabulation, spatial OLAP requires summaries in cartographic forms. In addition, standards for aggregation operators on geometric types have not emerged (Shekhar and Chawla 2003), making spatial data warehousing more difficult. Active involvement of domain experts is therefore needed to mature the development of this field. In most cases, background knowledge in the form of concept hierarchies is vital to assist these operators. For example, Shekhar et al. (2001) proposed the use of a map cube as a spatial analog of the data cube, aimed at generating an album of maps corresponding to all possible aspatial and spatial summaries of the data based on a specified spatial aggregation hierarchy. This is done by including standard summaries and cross-tabulations as well as spatial summaries at different levels of aggregation with pointers to the corresponding spatial objects along with geographic visualization.

Another central component of the GKD process is spatial data mining, which focuses on the development of theory, methodology, and practice for the extraction of useful

information and knowledge from massive and complex spatial databases. Conventional data mining approaches such as clustering, classification, association mining, and outlier analysis need advancements to handle the unique characteristics represented by the spatial components of spatial data. Deeply rooted in both traditional spatial analysis and various data mining approaches, spatial data mining has become an active research domain (Mennis and Guo 2009). Various types of spatial data mining techniques such as spatial classification and predication, spatial clustering, spatial outlier detection, and spatial association rule mining, have been developed.

First, spatial classification and prediction (e.g. (Koperski et al. 1998; Ester et al. 2001) basically map spatial objects into meaningful categories while considering the distance, direction or connectivity relationships and/or the morphology of these objects. Second, spatial clustering exploits spatial relationships among data objects in determining inherent groupings of the input data (Miller 2007). Han et al. (2001) suggested that many traditional methods, such as partitioning methods (i.e. k-means and the expectation-maximization (EM) method), hierarchical methods (i.e. top-down by splitting or bottom-up through aggregation), density-based methods, grid-based methods, model-based methods and constraints-based methods, can be adapted to spatial data. Third, aspatial outlier is defined as a spatially referenced object whose non-spatial attributes appear to be inconsistent with other objects within some spatial neighbourhood (Shekhar et al. 2003a). Computational strategies for detecting the outliers based on a single semantic attribute or spatial property such as size and shape has been proposed. Ng (2001) added distance-based measures to detect unusual paths traced by individual movement through a monitored environment in two-dimensional space. These measures are useful as they

allow the identification of unusual trajectories based on entry/exit points, speed and geometry and thus help in detecting unwanted behaviors such as theft.

This research particularly focuses on the use of spatial association rule mining techniques and assumes the availability of a given spatial data warehouse for mining. The geographic knowledge discovery process using spatial association rule mining is referred herein as spatial association rule (SAR) discovery. Association rule mining is regarded as one of the most popular mining techniques because it is efficient at finding frequent and meaningful relations, positive associations, and stochastic plus asymmetric patterns in large relational data warehouses. The concepts and recent advances in association rule (AR) mining are reviewed in the next session. A discussion of the challenges and achievement of spatial association rule (SAR) mining follows.

2.2 AR Mining: Concepts, Algorithms, and Recent Advances

Association rule mining has received significant attention since its introduction in 1993 by Agrawal et al. (1993) and is still regarded as one of the most popular approaches for pattern discovery in databases. It aims to identify association rules derived from sets of items that satisfy the predefined minimum support and confidence from a given database. The problem is formally stated as follows. Let $I = \{i_1, i_2, \dots, i_m\}$ be a set of items and D be a set of transactions, where each transaction T contains a set of items such that $T \subseteq I$. Associated with each transaction is a unique identifier; we say that a transaction T contains X , a set of some items in I if $X \subseteq T$. An association rule is an implication of the form $X \rightarrow Y$, where $X \subseteq I$, $Y \subseteq I$, and $X \cap Y \neq \emptyset$. X is called precedent and Y is often called antecedent. The rule $X \rightarrow Y$ holds in the transaction set D with confidence c if $c\%$ of transactions in D that contain X also contain Y . The rule $X \rightarrow Y$ has

support s in the transaction set D if $s\%$ of transactions in D contain $X \cup Y$. It is important that support is not confused with confidence. While confidence is a measure of the rule's strength, support corresponds to statistical significance. Originally, association rule mining was designed for market-basket data analysis where it aims to find rules like “A customer who buys products X_1 and X_2 , also buys product Y with probability $c\%$ ”.

Transaction-based association rule mining can be applied to non-transaction datasets such as object relational databases with a slight modification of the definition in which objects are regarded as transactions and their attributes, expressed in the form of predicates, are items. In this context, an association rule is defined as a dependence rule in the form of $X \rightarrow Y$ which can be explained as, “if a pattern X appears in the dataset, then the pattern Y tends to hold in the same dataset”, where X and Y are a set of one or more attributes.

The problem of association rule mining is usually decomposed into two sub-problems. The first is to find those itemsets whose occurrences exceed a predefined threshold in the database; those itemsets are called frequent or large itemsets. The second is to generate association rules from those large itemsets with the constraints of minimal confidence. Since the second sub-problem is quite straightforward, research has mostly focused on the first sub-problem, i.e. finding frequent itemsets. The solution to this problem involves two steps: candidate generation and check for frequent items. The Apriori-based algorithm is the most popular algorithm (Agrawal et al. 1993) because of its efficiency during the candidate generation process with pruning techniques to avoid measuring certain itemsets, while guaranteeing completeness (Kotsiantis and Kanellopoulos 2006). Basically, the candidate generation is performed with multiple

passes over the dataset. In a pass, the algorithm counts candidate itemsets by using only the itemsets found frequently in the previous pass – without considering the transactions in the database. The basic intuition is that any subset of a frequent itemset must also be frequent. The pseudo Apriori-like algorithm is summarized in Figure 2.

```

 $F_k = \{\text{frequent itemsets}\}; k = 1;$ 
While  $\text{card}(F_k) \geq 1$  do begin
     $C_{k+1} = \text{new candidates generated from } F_k;$ 
    For each transaction  $t$  in the database do
        Increment the count of all candidates in  $C_{k+1}$  that are contained in  $t$ ;
     $F_{k+1} = \text{candidates in } C_{k+1} \text{ with at least minimum support}$ 
     $k = k+1$ 
End
Answer =  $\bigcup \{F_k : k \geq 1\}$ 

```

Figure 2: The pseudo Apriori-like algorithm

The complexity and computational expense of the candidate generation process, as well as the requirement for multiple scans of the database, are bottlenecks of Apriori-based algorithms. Literature shows significant efforts to increase the efficiency of this algorithm either (a) by reducing the number of passes over the database, (b) by sampling the database, (c) through parallelization, or (d) by adding extra constraints on the structure of patterns.

Approaches aimed at reducing the number of passes over the database frequently utilize tree structures, such as Frequent Pattern (FP)-Tree (Han and Pei 2000) or TreeProjection (Agarwal et al. 2000) to store frequent items. FP-tree is an extended prefixed tree structure used to store information about frequent patterns of the database, consisting of frequent length-1 items stored as nodes. The tree nodes are arranged so that

nodes with more frequent occurrences will have better chances of sharing nodes than ones with less frequent occurrences. As FP-tree contains only frequent items and other irrelevant information are pruned, it is a compressed representation of the original database. The frequent itemsets are generated from this tree after only two passes over the database and without any candidate generation process. This approach, however, is not efficient for an interactive system with varying support and confidence thresholds or incremental databases. In these cases, the tree has to be reconstructed, i.e. the whole mining process need to be repeated. Conversely, TreeProjection uses a lexicographical tree to represent frequent mined patterns. This helps to reduce the support counting space, facilitates the management and counting of candidates, as well as provides the flexibility of picking an efficient strategy during the tree generation and transaction projection phrases. Other than tree representation, some authors suggest the use of logical operations (Wang and Tjortjis 2004) or binary based matrix (Yuan and Huang 2005) to store frequent set information with the need for scanning the database only once.

Database sampling can also be used to reduce the size of the dataset and increase the efficiency of mining algorithms. Ideally, the frequent itemsets extracted from a randomly drawn sample of transactions from the database represent a good approximation to the actual frequent itemsets extracted from entire database. For example, the algorithm proposed by Toivonen (1996) randomly picks a sample of the dataset and builds a candidate set of frequent itemsets containing all the frequent itemsets with a probability that depends on the sample size. However, the sample does not guarantee that all itemsets in the candidate set are frequent; neither to include the frequent itemsets found as if the whole dataset was processed. Nevertheless, the set of candidates allows the algorithm to

efficiently identify the set of frequent itemsets with at most two passes on the entire dataset. As the accuracy of sample-based association rule mining depends very much on sample size, some researchers have advocated progressive sampling as a better solution. Fundamentally, progressive sampling involves the analysis of increasingly larger samples until the observed improvement of a certain measure of the accuracy of the sample with respect to the mining task falls below a specified threshold. For example, Parthasarathy (2002) advanced this method based on a novel measure of self-similarity of associations across progressive samples, and a refinement technique based on equivalence classes to identify a proper sample size. Another example includes the work by Chuang et al. (2005), which aims to identify an appropriate sample size for mining association rules based on Sampling Error Estimation (SEE).

Parallel computing is another important trend to deal with the huge volume of data and associated processing. This requires a partition of the database among the processors. From the perspectives of hardware which supports parallelism, parallel computing systems could be classified, but not mutually excluded, into (1) multicore computing, (2) symmetric multiprocessing, and (3) distributed computing (Roosta 2000; Asanovic et al. 2006; Barney 2013). The first type, multicore computing, refers to a single computer having multiple computing units (i.e. cores) built on the the same chip. The second type, symmetric multiprocessing, refers to a computing system containing multiple processors that share resources and use bus for communication. In this type of systems, memory space and attached disks are shared by processors. Processors communicate through shared variables in memory and are capable of accessing any memory location. Synchronization is required to coordinate processes. The programming architecture on

these machines is challenging due to requirements to ensure efficiency and avoid failure. Particularly, memory latency is a significant bottle neck for this type of system and thus requiring data locality optimization (i.e., having as much as possible of the data local to the processor cache) and elimination of false sharing (i.e., the problem where coherence is maintained even when memory locations are not shared). The third type, a distributed computing facility or a distributed memory multiprocessor, refers to computer system composing distributed memory multiprocessors connected via a network. In such system, each processor has private memory and local disk to handle the distributed portions of a same task, and communicates with other processors only via passing message. Scalability is one of the biggest advantages. The most current types of distributed computing include cluster computing, massive parallel processing (MPP) and grid computing differentiated by types of connectivity. A cluster composed of many computers connected using a network created by commodity hardware. A massive parallel processor is similar to a cluster, but have specialized connecting network rather than hardware based connectivity. In contrast with clusters and MPPS, grid computing as well uses networked computers but communicate over the Internet. The most current and fastest parallel computing systems today employ hybrid distributed-shared memory, i.e. both shared and distributed memory, architecture. While this type of system requires programming complexity, it has been indicated to be the one to prevail in the future of high end computing due to the combined capability to handle computational expensive processing as well as scalability (Barney 2013). Regarding parallel computing for association rule mining in particular, early research has been focused on distributed memory systems (Agrawal and Shafer 1996; Cheung et al. 1996; Han et al. 1997a; Zaki

et al. 1997). Examples of algorithms developed for this type include FDM - Fast Distributed Mining (Cheung et al. 1996), which were later advanced into FPM - Fast Parallel Mining (Cheung and Xiao 1998) with a more optimal communication scheme. In these cases, the database scan is performed independently on the local partition and support counts of the local frequent candidate sets must be exchanged among all the sites to find global frequent sets. Techniques such as Principal Components Analysis could be used to improve the data distribution prior to parallel mining as suggested with the Data Allocation Algorithm (Manning and Keane 2001). Recent trend in parallel computing for association rule mining has moved toward symmetric multiprocessing systems (SMPs), often called shared-everything systems due to the capability of delivering high performance at economical cost. More detailed discussions and proposed solutions for this matter can be found in (Parthasarathy et al. 2001).

Adding constraints on the structure of patterns mined is another way to speed up the mining process. This is done in three ways: (1) dataset filtering by restricting the source dataset to objects that can possibly contain patterns satisfying the constraints (Wojciechowski and Zakrzewicz 2002); (2) pattern filtering via an integration of pattern constraints into the actual mining process in order to generate only patterns satisfying the constraints (Do et al. 2003); or (3) post-processing by filtering out patterns that do not satisfy user-specified pattern constraints after the actual discovery process.

As association rule mining has found its way into many application domains, various advances in algorithms are found, involving the concepts of quantitative association rule (Srikant and Agrawal 1996), multi-level (taxonomies) association rule – also known as generalized association rule (Srikant and Agrawal 1995), fuzzy association

rule (Ladner et al. 2003), sequential temporal patterns (Agrawal and Srikant 1995; Verma et al. 2005), and spatial association rule. As the objective of this research is to advance association rule mining particularly for geospatial analysis, the following sections are dedicated to review the definitions and recent developments in spatial association rule mining.

2.3 SAR Mining: Challenges and Achievements

2.3.1 Definition of SAR

The adaptation in definition of spatial association rules from mainstream association rules is simply by the inclusion of at least one predicate that is spatially defined (Koperski and Han 1995). Spatial predicates in spatial association rule mining are used to express spatial information, i.e. spatial components found in the dataset, while non-spatial predicates are for semantic information. An example of non-spatial predicates is $\text{is-city}(X)$ which is used to express the true or false property of semantic information, i.e. whether the object X is a city or not. An example of spatial predicate is $\text{close-to}(X, Y)$, which would be used to express the fact that objects X and Y are close to each other. A spatial association rule is then defined as a rule indicating an association relationship among a set of spatial and possibly non-spatial predicates (Koperski and Han 1995). A strong rule indicates a pattern which has relatively frequent occurrences in the database and thus suggests a strong implication relationship between (or among) predicates of the rule. Formally, a spatial association rule mining problem can be stated as following: Let D be a spatial database that contains spatial objects O for studying and let P be a set of all possible predicates, both non-spatial and spatial, that could be derived from D . Each object O has a unique identifier and an object O possesses X , a set of some predicates in

P , if $X \in P$ and X represent semantic or spatial properties of O . A spatial association rule is an implication of the form $X \rightarrow Y$, where $X \in P$, $Y \in P$, $X \cap Y \neq \emptyset$, and $\exists(x \in X \text{ or } y \in Y) | x, y \text{ are spatial predicate}$). The rule $X \rightarrow Y$ holds in D with confidence c if $c\%$ of objects in D that contain X also contain Y . The rule $X \rightarrow Y$ has support s in D if $s\%$ of objects in D contain $X \cup Y$. For example, a rule like “is-large-city(X) and within(X ,California) \rightarrow close-to-water(X)” is a spatial association rule. This rule states that if X is a large city and X is within the state of California in the US then X is close to the sea.

Generally, the problem of spatial association rules can be decomposed into three steps: (a) derive non-spatial/spatial predicates, (b) find all frequent sets of predicates, and (c) generate strong association rules. Although the deployment of association rule mining for spatial datasets is theoretically straightforward, the unique characteristics of spatial data discussed earlier engender challenges for spatial association rule mining. As spatial components are presented using predicates, spatial predication, i.e. the process of identifying, formulating, and capturing spatial components using spatial predicates, is crucial for spatial association rule mining.

2.3.2 Existing SAR Mining and Discovery Algorithms and Frameworks

Research in spatial data mining has received some attention in recent years, evidenced by the development of a handful of algorithms and frameworks which are to some extent successful in dealing with spatial data.

One approach to SAR mining follows a framework designed to work with singular relational database (i.e. table) while facilitating procedures for spatial predication. Examples of this framework include GeoMiner, which is an extension of a relational data

mining system called DBMiner (Koperski and Han 1995; Han et al. 1997b), and Weka-GDPM (Bogorny et al. 2006a; Bogorny 2006b; Bogorny 2006c; Bogorny et al. 2008a; Bogorny et al. 2008b; Bogorny et al. 2010), an extension of a data mining software named Weka (Bouckaert et al. 2011). Under this type of framework, spatial predicates generated on the basis of spatial attributes and spatial relationships of the participating features are stored in the mining table columns in addition to the ones of semantic attributes. The Apriori-based algorithm is often utilized for frequent itemset discovery.

A second type of SAR mining framework argues against the use of singular relational database for the reason that mining rules at the level of a single concept is not very effective for spatial data due to the problem of granularity. For example, Malerba (2008) states that concept hierarchies in spatial data are a valuable kind of domain knowledge to be exploited during pattern discovery. This is because it is more likely to discover interesting rules at low concept levels than at higher ones, while large support is more likely to exist at a high concept level rather than at low ones. It is then suggested that multi-relational algorithms, which operate on data scattered through multiple tables (relations) of a multiple relational database, are more promising for spatial data mining problems. The supporting argument for this is that a multi-relational setting can deal with the heterogeneity of spatial objects, can distinguish their different role (reference or task-relevant), can naturally represent a large variety of spatial relationships among objects, and can also accommodate different forms of spatial autocorrelation (Malerba 2008).

The literature reports on a couple of systems which perform spatial association mining according to this second approach. The first one is the SPIN! platform (May and Savinov 2003), which assumes an object-relational data representation and offers facilities for

multi-relational sub-group discovery and multi-relational association rule discovery. Subgroup discovery (Klosgen and May 2002) takes advantage of a tight integration of the data mining algorithm with the database environment. Spatial relationships and attributes are then dynamically derived by exploiting spatial DBMS extension facilities (e.g., packages, cartridges, or extenders) and used to guide the subgroup discovery. This means that, since the number of spatial relationships between data layers can be very large and many of them might be unnecessarily extracted, the subgroup discovery approach dynamically performs spatial joins only for the part of the hypothesis space that is really explored during the search by a data mining algorithm. The second one is SPADA (Spatial Pattern Discovery Algorithm) (Lisi and Malerba 2002). SPADA was initially designed according to the theoretical framework of inductive databases, which can be perceived as a deductive database (e.g. spatial database – SDB) with an integrated inductive component (e.g. data mining engine - DM Engine) as shown in Figure 3. The inductive component, DM Engine, is an ILP (inductive logic programming) module that supports the processing of spatial association mining queries. The integration of the ILP module requires data and patterns to be represented in a logical language and this, beneficially, allows the specification of rich domain knowledge such as spatial hierarchies, spatial constraints, and rules for spatial qualitative reasoning (Ceri et al. 1991). The DM Engine aims to discover patterns according to an increasing order of description granularity, i.e. from coarser-grained to finer-grained. Frequent patterns are generated by the level-wise method (Mannila and Toivonen 1997), which is based on a breadth-first approach searching the lattice spanned by a generality between patterns. The DM Engine returns as many *.pat and *.rul files at the number of description granularity

levels; each rule reports frequent patterns and strong rules either in text format or in XML format.

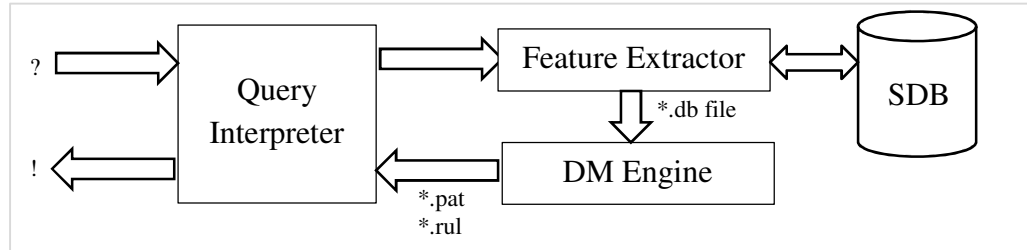


Figure 3: Software architecture of SPADA (Lisi and Malerba 2002)

Within the SPADA system, a feature extraction module is implemented to pre-compute spatial relationships, which are then converted into Prolog facts used by the ILP-based DM Engine system. Spatial index structures, such as R-trees (Guttman 1984), are used to speed up the processing of spatial joins. SPADA makes use of Oracle Spatial databases, which support the vector format and the 9-intersection model for the computation of topological relations. The SPADA algorithm is then used in the development of ARES (Association Rules Extractor from Spatial data) (Appice et al. 2005) for the specific purpose of mining association rules and INGENS 2.0 (Malerba et al. 2010) to mine both association and classification rules. ARES, which requires access to Spatial Oracle, is freely available to the research community.

In addition, the literature also contains a number of contributions that handle other spatial data properties such as heterogeneity and fuzziness. To cope with spatial heterogeneity of the event being analyzed using SAR mining, Li (2008) used a moving window integrated with the Apriori-based algorithm. The approach however remains limited for discovering association between two variables only. Fundamentally, a window is used to move over the region of interest during the mining process. Associated with

each window location is an association rule of the two variables with a certain degree of support and confidence. These estimated supports and confidences are then used to generate the so-called three-dimensional association intensity for the whole study region, which allows a visual extraction of sub-regions having uniform support and confidence pattern. The use of different window sizes and shapes are suggested by the author for sensitivity tests.

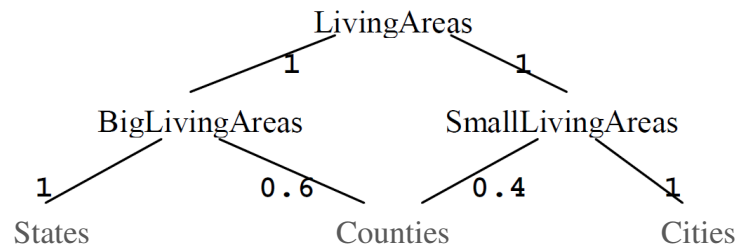


Figure 4: An example of fuzzy concept hierarchy (adapted from Kacar and Cicekli 2002)

The literature recognizes that fuzziness may be an issue in spatial association rules because linguistic expressions are used in both aspatial and spatial predicates. For semantic information, examples involve the classification of cases into categories such as young and old (for age), big and small (e.g. for city size), or high and low (e.g. number of crime incidents), expensive and cheap (e.g. for housing property), to name a few (e.g. Buczak and Gifford 2010). For spatial cases, according to (Kacar and Cicekli 2002), fuzziness exists in spatial concept hierarchy and spatial relation hierarchy in the sense that an item could partially belong to more than one parenting items. For instance, in the concept hierarchy depicted in Figure 4, 'counties' can be regarded as 'big living area' or 'small living area' with membership function value of 0.6 and 0.4, respectively. Similarly, in the case of the spatial relation hierarchy shown in Figure 5, the 'overlap'

relationship belongs to ‘intersect’ and ‘inside’ relationship with equal membership function values of 0.5.

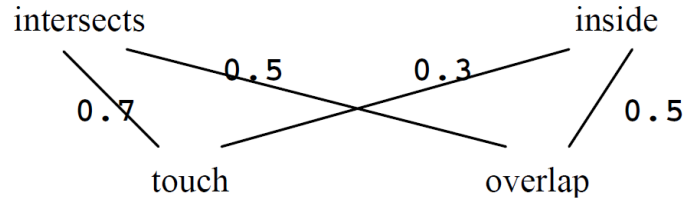


Figure 5: An example of fuzzy spatial relationship hierarchy (adapted from Kacar and Cicekli 2002)

Also dealing with fuzziness in spatial predicates, Laube et al. (2008) exclusively focuses on spatial proximity relations between spatial entities and investigates the approaches to define suitable distance measures between various types of spatial objects (i.e. point, linear, areal) to compute these distance measures efficiently, and to map these distance measures to scores in the range $[0, 1]$ for use in fuzzy spatial association rules. The authors in addition presented a conceptual framework to calculate the quality measures of spatial fuzzy association rules in term of support and confidence. By this method, a product t-norm is used over the fuzzy membership function values of the associating predicates. This can be demonstrated by considering a simple fuzzy spatial association rule in which both the antecedent (i.e. left hand side of the rule) and the consequent (i.e. the right hand side of the rule) consist of a single predicate, such as “If a house is close to the sea, then it is expensive”. Scores $s_{Ant}(H)$ and $s_{Cons}(H)$ in the range $[0, 1]$ are used to capture to what extent a house H is close to the sea and to what extent a house is expensive, respectively. The support and confidence of this rule, referred as spatial support and spatial confidence by the authors is formulated as:

$$\begin{aligned}
\text{spatial support} &= \sum_H s_{\text{Ant}}(H) \times s_{\text{Cons}}(H) \\
\text{spatial confidence} &= \frac{\sum_H s_{\text{Ant}}(H) \times s_{\text{Cons}}(H)}{\sum_H s_{\text{Ant}}(H)}
\end{aligned}$$

where the sum is over all houses H . A similar formulation can also be found in (Ladner et al. 2003). In cases where the antecedent and/or the consequent of a SAR comprise several predicates with an AND relationship, the product t-norm is again used to estimate the overall score of the antecedent and/or the consequent. For example, considering a rule “If a house is close to the sea AND close to a big city, then it costs at least \$800,000.” If the score for “close to the sea” is 1 and that for “close to a big city” is 0.5 then the overall score for the antecedent is $1 \times 0.5 = 0.5$. The support and confidence of the rule is then estimated using the overall scores for antecedent and the consequent as the previous case of having single predicate on each side of the rule. If the antecedent and/or the consequent of a SAR consist of several predicates with an OR relationship among them, a t-conorm which is similar to a t-norm except that it has 0 as identity element is suggested to calculate the overall score of the of the antecedent and/or the consequent. An example of such is the Einstein sum defined as $\frac{s_1+s_2}{1+s_1s_2}$ for two scores s_1 and s_2 . For example, considering a rule “If a house is close to a highway or close to an airport, then it has good sound insulation.” and the score for being close to highway is s_1 , and that for being close to airport is s_2 . In cases where a house is already very close to a highway ($s_1 = 1$), then its proximity to an airport is irrelevant for the support of the rule. The overall score for the antecedent is $\frac{1+s_2}{1+1*s_2}$, which equals to 1 regardless of the s_2 value. But if the house is somewhat close to a highway and somewhat close to an airport (both with score 0.5),

then the overall score for the antecedent is $\frac{0.5+0.5}{1+0.5*0.5} = 0.8$. The support and confidence of the rule is then estimated using the overall scores for the antecedent and for the consequent in a way similar to having single predicate on each side of the rule.

A shortfall of the SAR mining and discovery frameworks discussed above is that they neglect complex and dynamic spatial dependence structures of the phenomena under study during the process of spatial predication. In addition, these platforms have disregarded the importance of mined result evaluation, which is a crucial procedure of knowledge discovery. A discussion on these two remaining challenges is presented in the following sections.

2.3.3 Spatial Predication – The First Remaining Challenge

Regarding the first challenge, it is argued here that spatial effects under the form of spatial dependence, which have remained a central concern of quantitative geographers, regional scientists, as well as spatial econometricians for the last forty years or so, should also be addressed in the process of spatial association rule mining, particularly during the formation of spatial predicates.

It seems logical to begin the argument by recapitulating the necessity to handle spatial dependence (or spatial autocorrelation, interchangeably) in traditional spatial statistical analysis and modeling, before recasting the direct relevance of this issue to the case of spatial association rules. By definition, spatial dependence, or spatial autocorrelation, originally explicated in (Cliff and Ord 1970) is often taken to mean the lack of independence among observations. Hubert et al. (1981, p. 224) provided a formal definition of spatial dependence as: “Given a set S containing n geographical units, spatial autocorrelation refers to the relationship between some variable observed in each

of the n localities and a measure of geographical proximity defined for all $n(n-1)$ pairs chosen from n .''' Accordingly, values for the same attribute measured at locations that are near to one another tend to be similar, and tend to be more similar than values separated by larger distances (Haining 2003). The awareness of the problems caused by spatial dependence and their effects on the validity of statistical methods is dated as far back as Student (1914). The development of spatial statistics, however, remained naïve until the derivation of the first formal indices to detect the presence of spatial dependence by (e.g. Moran 1948; Geary 1954; Dacey 1968). Regional scientists and geographers only gained real exposure to spatial statistics in the late 1960s and 1970s (Cliff and Ord 1969; Cliff and Ord 1973; Ord 1975; Sen 1976; Haining 1978); nonetheless they came to an immediate realization of their significance. Evidence of this is a substantial quantity of research focused on constructing several techniques to test for spatial dependence, as well as on understanding characteristics of this particular matter. In term of testing, various spatial autocorrelation statistics and extensions to multivariate analysis have been constructed (e.g. (Royalty et al. 1975; Sen and Soot 1977; Hubert et al. 1981; Hubert 1985; Wartenberg 1985). Moran's I , Geary's C (Cliff and Ord 1973), Getis-Ord's G (Getis and Ord 1992), Ripley's K (Ripley 1977) are now common global statistical indicators. Indicators for local scale analysis include Getis-Ord's G_j and G_j^* (Ord and Getis 1995), Anselin's I_i and c_i (local indicators of spatial association (LISA) (Anselin 1995), and Ord-Getis' O (taking into account global autocorrelation) (Ord and Getis 2001).

Effects of spatial dependence on the estimation, identification, and model specification of spatial process models have also been extensively studied (e.g. Cliff and

Ord 1969; Cliff and Ord 1973; Haining 1977, 1979; Anselin 1986a, 1986b; Haining 1986; Anselin 1988; Anselin and Griffith 1988; Anselin 1990; Getis 2008). Under the impacts of spatial dependence, the popular regression analysis technique for cross-sectional data experiences situations in which the dependent variable at one location may be functionally related to its own value at some other locations. If left untreated, spatial dependence will cause bias and spatial errors in regression residuals which invalidate the interpretation of standard hypothesis tests and estimates (Cliff and Ord 1973; Anselin 1988). From a modeling perspective, spatial dependence is considered as the existence of a functional relationship between what happens at one point in space and what happens elsewhere. Thus, failure to treat spatial effects when necessary will misspecify the models (Anselin 1988). At a more general level, Getis (2008) emphasized the significance of the spatial autocorrelation concept as it “provides tests on model misspecification; determines the strength of the spatial effects on any variable in the model; allows for tests on assumptions of spatial stationarity and spatial heterogeneity; finds the possible dependent relationship that a realization of a variable may have on other realizations; identifies the role that distance decay or spatial interaction might have on any spatial autoregressive model; helps to recognize the influence that the geometry of spatial units under study might have on the realizations of a variable; allows us to identify the strength of associations among realizations of a variable between spatial units; gives us the means to test hypotheses about spatial relationships; gives us the opportunity to weigh the importance of temporal effects; provides a focus on a spatial unit to better understand the effect that it might have on other units and vice versa (‘local spatial autocorrelation’);

helps in the study of outliers”. Thus, “no other concept in empirical spatial research is as central to model building as is spatial autocorrelation” (Getis 2008).

This so-called “fundamental element of all spatial models” Getis (2008) is directly relevant, with no exception, to the problem of mining spatial association rules. Explicitly, spatial autoregressive models use a coefficient-based component to express the spatial dependence (i.e. $\rho W y$ in spatial lag model: $y = \rho W y + \beta x + u$, or $\lambda W \varepsilon$ in spatial error model: $y = \beta x + \lambda W \varepsilon + u$). For spatial association rules, let us consider a given rule: if “block-group A is next to block-groups of high crime” then “block-group A has high crime”. In this case, the spatial spillover effects of crime is captured at the block group level, after providing specifications in defining predicates “next-to” and “high crime”. Indeed, spatial association rules are one form of spatial modeling using rule-based linguistic expressions to convey associative implication regarding aspatial and spatial characteristics of analyzed features (or variables) as well as spatial effects and spatial interactions among them. As in regression analysis, spatial autocorrelation violates the assumption of independent transactions in mining association rules. Failing to account for this will result in the omission of important functional implications due to spatial effects.

The existing SAR literature has up to now failed to address this matter. Spatial association rules are distinguished from non-spatial ones by the inclusion of spatial predicates. By using linguistic expressions, spatial predicates allow flexible expressions related to explicit spatial relations of objects in terms of distance (e.g. close-to), direction (e.g. north-of), and topology (e.g. adjacent-to) but also to implicit spatial dependencies, i.e. spatial autocorrelation, or more generally the spatial dependence structure imbedded in the phenomena under study. However, the dynamics and complexity of spatial

components captured within these spatial predicates are typically overlooked. To the best of our knowledge, there is yet no mechanism that generates predicates to capture spatial dependencies.

The complete disregard for modeling spatial dependence structures and integrating complex spatial components in SAR mining could be ascribed to the limitation in geographical domain knowledge possessed by computer science mining experts who have been so far the frontrunners in the field of SAR mining. Spatial predicates are commonly limited to spatial relations rather than spatial dependencies (Malerba et al. 2002; Appice et al. 2003; Malerba et al. 2003; Bogorny 2006a). Due to the large number of spatial relations in large databases, much research has focused on developing algorithms to efficiently extract them. For example, Koperski and Han (1995), Koperski and Han (1996), and Koperski (1999), in GeoMiner, proposed a top-down progressive refinement method towards spatial query results with which coarse spatial approximations are calculated first, and then, more precise spatial relationships are later computed. Other examples include the Spatial Pattern Discovery Algorithm (SPADA) (Lisi and Malerba 2002) and SPIN! (May and Savinov 2003). SPADA pre-computes distance, direction and topological relations and materializes (i.e., stores) them into some database relations (called neighborhood indices), which are then used by data mining algorithms to efficiently retrieve all neighbors (with respect to some spatial relation) of a given spatial object. SPIN!, on the other hand, partitions the database into subgroups to reduce the number of spatial relations to be computed.

Some research has attempted to incorporate geographical knowledge in the process of spatial association rule mining, although in a fairly simple fashion. For instance,

(Bogorny et al. 2005; Bogorny et al. 2006a; Bogorny et al. 2006b; Bogorny et al. 2008a; Bogorny et al. 2008b; Bogorny et al. 2010) refer to the concept of “well-known geographic dependence” defined to be the obvious geographical relations in the form of $A \rightarrow B$ with 100% support, such as $\text{is_an(island)} \rightarrow \text{within(water)}$, $\text{is_a(bridge)} \rightarrow \text{cross(water)}$, or $\text{is_a(gasStation)} \rightarrow \text{intersects(street)}$. It is advised that these well-known dependencies can be identified using geographical domain knowledge base in order to set up constraints to prune the input space of spatial predication. This not only helps to reduce the input dimension, to speed up the computation process of spatial joins to produce a more efficient set of spatial predicates, but also to avoid the generation of a large number of patterns and rules without novel, useful and interesting knowledge. It should be noted that this so-called “well-known geographic dependence” concept merely involves spatial relations among objects and is different from the concepts of spatial dependence and spatial autocorrelation that we discussed earlier.

There have been a few attempts at using association rule mining particularly toward geospatial problems; however the treatment of spatial dependence has remained completely neglected. For example, Mennis and Liu (2005) were interested in using association rule mining to explore the relationships among a set of variables that characterize socioeconomic and land cover change in the Denver, Colorado region from 1970 – 1990. The associative variables were the change in percent of minority, change in percent of population living below the poverty line, and density of developed land. No spatial components have been considered in this study. Jung and Sun (2006) used spatial association rule mining to discover factors associated to location choices of convenience stores in Taipei City. Apart from various aspatial demographic-related attributes

including population, sex, age, education level, and job types, spatial attributes such as distances to train stations, to gas stations, to police stations, to post offices, to restaurants, to hospitals, to banks, and to roads were also considered. The mined rules are then used in a decision support system for locating new convenience stores in the city. Lee and Phillips (2008) proposed a framework to detect multivariate associations used in ArcGIS based on a given areal base map. With this approach, the feature datasets are aggregated onto the areal base map and categorized into several groupings. Association rules are then mined out of the attribute table linked with the areal base map. The framework is applied to discover associations to crime cases for Brisbane, Australia. In order to generate spatial predicates, the study includes distances to various geospatial features namely parks, lakes, highways, rivers, schools, hospitals, airports, highways, transit stations, police stations, and post offices. These studies are confined by limitations in handling components of spatial processes involved in the phenomena under study such as spatial dependence structures and spillover effects of events, as well as spatial interactions among participating features. Lee and Estivill-Castro (2011) necessitated the use of spatial clustering (Lee and Estivill-Castro 2006) before mining for association rules in a so-called horizontal-view association mining approach. This technique works on the principle of overlaying GIS-based clustered layers in order to identify overlapping areas and, thus, generate frequent association rules. The study however emphasizes the use of spatial clustering for data classification for rule mining rather than concerns related to spatial dynamic effects.

The comprehensive spatial association rule mining framework envisioned for this study endorses methodologies to (1) robustly identify complex structures of spatial

dependence, and (2) incorporate it to the process of spatial predication, so as to subsequently allow the discovery of associations governed by spatial dependence, if any. The problem of identifying the correct specification of the spatial dependence structure itself has remained until recently a challenge drawing substantial attention and debate among spatial analysts and spatial econometricians. In addition, the construction of the spatial predicates that fully and accurately integrate the spatial dependence effects with well-defined semantics remains an unresolved research question in SAR mining. Furthermore, careful examination of the effects of ignoring spatial dependence as well as of structural misspecifications in spatial association rule mining and discovery are beneficial to further developments in SAR mining applications.

2.3.4 Evaluation and Visualization – The Second Remaining Challenge

Another crucial component of the knowledge discovery process is the evaluation of the mined results. In association rule mining, this is especially critical when a large number of patterns are generated, particularly when dealing with a large dataset, or having small support and confidence thresholds. As the size and dimensionality of the database increase, the generation of millions of patterns is not uncommon, many of which may be uninteresting. Challenges remain not only in developing evaluation approaches that are highly automated, but also in establishing a well-accepted set of criteria for evaluating the quality of association patterns. These problems are often intensified when dealing with spatial databases. First, large and complex spatial datasets often involve many predicates to represent semantic attributes and spatial components, so that a substantial number of frequent itemsets and rules is produced. Second, the complex nature of spatial data involving dynamic representation forms and implicitly embedded

spatial patterns challenges automated and objective-based evaluation schemes. Thus, successful evaluation schemes for spatial association rule mining should facilitate more human-centered interactivity with subjective evaluation criteria constructed based on domain knowledge. Furthermore, given the increasing availability of “big” data sets, this can be fruitfully accomplished with a visual analytics approach.

Visual analytics in general is “the science of analytical reasoning facilitated by interactive visual interfaces” (Thomas and Cook 2005). Keim et al. (2008) defined visual analytics as an integral approach combining visualization, human factors and data analysis with the goal of gaining insight into a large information space; it combines automatic analysis methods with human background knowledge and intuition. Concerning visualization, the techniques found in this field often involve computer graphics, visualization metaphors and methods, information and scientific data visualization, visual perception, cognitive psychology, diagrammatic reasoning, 3D virtual reality systems, multimedia and design computing, and virtual environment (Simoff et al. 2008c). To deal with a vast amount of data from heterogeneous sources, visual analytics often combine the strengths of machines with those of humans. Methods from data mining and knowledge discovery in database, statistics and mathematics are often found to be the driving force for automation while human knowledge and analytical thinking are used to perceive, relate, and conclude. For the purpose of data mining, visual analytics may be referred to as visual data mining (Simoff et al. 2008b). By combining the respective strengths of humans and machines, decision makers can focus their full cognitive and perceptual capabilities on the analytical process, while applying advanced computational capabilities to augment the discovery process (Keim et al. 2008).

Geovisualization can be considered to be quite similar to visual analytics but with an emphasis on the geospatial domain. MacEachren and Kraak (2001) related geovisualization to the development of theories and methods which facilitate knowledge construction through visual exploration and analysis of geospatial data and the implementation of visual tools for subsequent knowledge retrieval, synthesis, communication, and use. Geovisualization is distinguishable from cartography, which is perceived as traditional visualization techniques in the spatial domain focusing on the design and use of maps for information communication and public consumption. On the other hand, geovisualization often emphasizes the development of highly interactive maps and associated tools for data exploration, hypothesis generation and knowledge construction (MacEachren 1994; MacEachren and Kraak 1997) and has very close relations with exploratory data analysis (EDA) and exploratory spatial data analysis (ESDA) (Tukey 1977; Bailey and Gatrell 1995; Anselin 1999). Development in geovisualization has, however, generally remained confined to data poor environments for linking statistical graphics and maps; it relies on human experts to interact with data, visually identify patterns, and formulate hypotheses or model. A few exceptions focus on addressing multiple perspectives and many variables simultaneously by coupling visualization with dimension reduction techniques such as multidimensional scaling (Young 1987), principle components analysis (PCA) (Abdi and Williams 2010), self-organizing maps (SOM) (Agarwal and Skupin 2008; Yan and Thill 2009), or other projection pursuit methods (Cook et al. 1995). Several approaches to multivariate mapping have also been developed, including specially designed symbols (Chernoff and Rizvi 1975; Zhang and Pazner 2004), multiple linked views (Monmonier 1989; Dykes

1998; MacEachren et al. 1999), and clustering-based approaches (Guo et al. 2003; Guo et al. 2005).

Nevertheless, data rich environments place further requirements on geovisualization so as to handle extraordinary large spatial datasets, and have thus evolved along with the trend of visual analytics. An amalgamation of automatic computation methods such as spatial data mining and geospatial analytics with effective designs and interactive strategies to facilitate the discovery process is indeed desirable (Mennis and Guo 2009). With some existing attempts (Andrienko and Andrienko 1999; Ward 2004; Guo et al. 2005), future research in geovisualization is heading towards what is referred herein as geo-visual analytics especially designed for geo-spatial domain.

Visual analytics for spatial association rule mining and geographic knowledge discovery in particular, possess certain features and modules from both geovisualization and visual data mining. Developing such system for SAR mining and discovery is a focus of this study and will be discussed in more detail in Chapter 5.

2.3.4.1 AR Visualization

Various studies exist on visualizing association rules and Bruzzese and Davino (2008) provides an good overview of the relevant techniques. The most common and traditional technique is table-based (Han and Kamber 2001). Each row of the table represents an association rule while the columns represent the items, the antecedents and consequents, the support, and the confidence of association rules. Figure 6 depicts a table representing association rules.

The second form of visualization is matrix-based (2D or 3D) (Wong et al. 1999; Hofmann and Wilhelm 2001). A two-dimensional association matrix positions the

antecedent and consequent items on separate axes of a square matrix. Customized icons are then used on the matrix tiles to connect the antecedent and the consequent of the corresponding association rules. Figure 7(a) shows an example of such visualization for rule $B \rightarrow C$. The height and color of columns are used to represent the properties of the association rules such as support and confidence. The 2D matrix-based visualization technique however breaks down when it is used to present many-to-one relationships. For example, in Figure 7(b), it is very difficult to distinguish if there is only one association rule ($A+B \rightarrow C$) or two ($A \rightarrow C$ and $B \rightarrow C$). The solution of grouping all the antecedent items of an association rule as one unit and plotting it against its consequent as shown in

	A	B	C	D	E	F	G
1	Antecedent items				Consequence	Confidence	Support
2	Breathes	Toothed			Backbone	1.00	0.47
3	Backbone	Milk	Toothed		Breathes	1.00	0.40
4	Breathes	Milk	Toothed		Backbone	1.00	0.40
5	0 Legs	Backbone			Tail	0.95	0.18
6	Backbone	Hair	Milk		Breathes	1.00	0.39
7	Breathes	Hair	Milk		Backbone	1.00	0.39
8	Backbone	Breathes	Hair	Toothed	Milk	1.00	0.38
9	0 Legs	Catsize			Tail	0.86	0.06
10	0 Legs	Predator			Eggs	0.76	0.13
11	Eggs	Fins	Predator	Toothed	Tail	1.00	0.09
12	Predator	Tail	Toothed	Venomous	Eggs	0.67	0.02
13	Tail				Toothed	0.69	0.51
14	>4 Legs	Eggs			Breathes	0.67	0.08
15	>4 Legs	Hairborne			Hair	0.67	0.04
16	0 Legs	Aquatic			Backbone	0.94	0.17
17	2 Legs	Aquatic	Eggs		Hairborne	0.83	0.05
18	2 Legs	Aquatic	Tail		Eggs	0.86	0.06

Figure 6: Table-based visualization of association rules

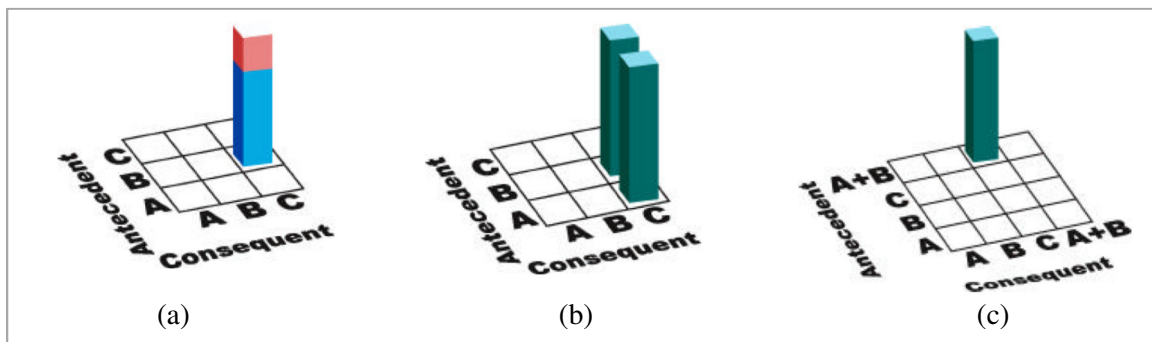


Figure 7: 2D-matrix based visualization of association rules

Figure 7(c) only works effectively with small antecedent sets. Another issue with the matrix-based approach is the occlusion problem, especially when multiple icons are used to depict different metadata values on the matrix tiles. A three-dimensional matrix shows rule-to-item rather than item-to-item relationships. Figure 8 depicts an example of such relationships. By this approach, the rows represent the items and the columns represent the item associations. Color is used to distinguish the antecedent and the consequent of the rules. Confidence and support levels of rules are shown by the corresponding bar charts in different scales at the far end of the matrix. Although a 3D matrix representation overcomes the problem of many-to-one relationships, the occlusion problem still remains for large sets of rules.

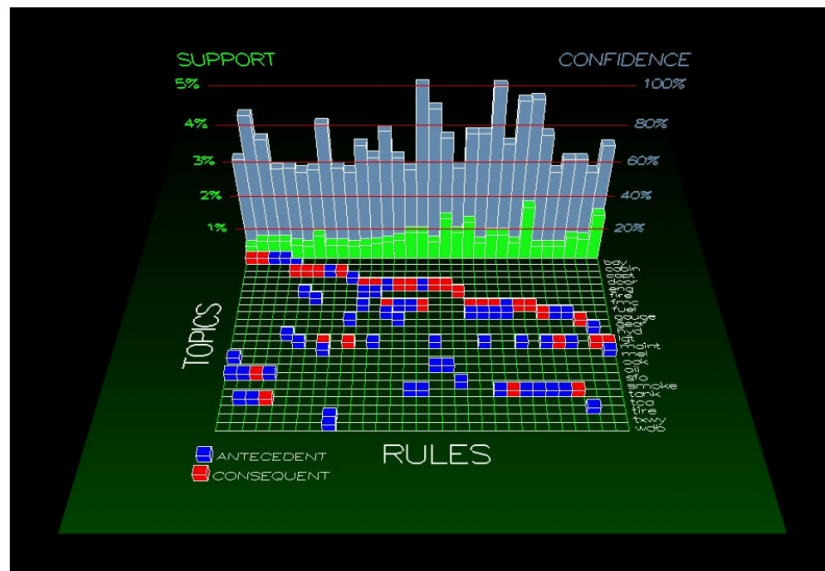


Figure 8: Three-dimension matrix visualization of association rules

The third form of association rule visualization is network-based, in which the nodes represent the items and edges represent the associations. One example of such type is using direct graph (Han and Kamber 2001), as shown in Figure 9. Different colors and

width of the arrows are used to represent properties of the rules such as the confidence and the support. When many rules with many items are represented, the direct graph is not easy to understand because of the superimposition of the edges with the nodes. Hetzler et al. (1998) proposed animating the edges and selectively showing associations of certain items with 3D rainbow arcs. However, this technique requires significant effort to turn on and off the item nodes. In addition, showing multiple metadata values such as confidence and support corresponding to the rules is not easy. Another example is shown in Figure 10 which uses a so-called association rule network (Statistica 2012). In this figure, a subset of 15 rules is displayed. The thickness of each line indicates the confidence of the rule while the size of the circles in the center indicates the support of each rule.

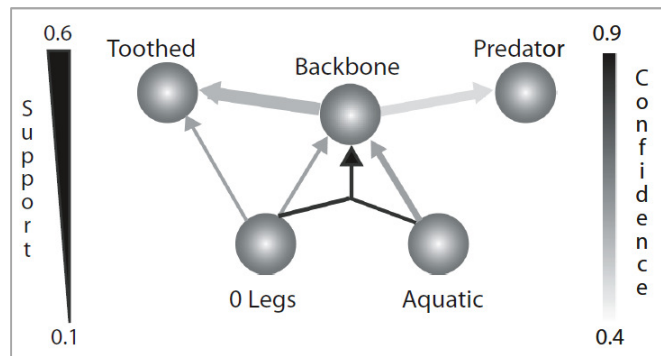


Figure 9: Direct graph visualization of association rules

A fourth form of association rule visualization is the TwoKey plot (Unwin et al. 2001), which represents the rules according to their confidence and support values. With this approach, rules are presented in a 2D space as shown in Figure 11, where the x-axis and the y-axis range from the minimum to the maximum values of the support and confidence, respectively. Colors are used to highlight the order of the rules. Selection and interactive features as well as linkage with other displays can then be used to explore

rules. The analysis of items associated with the displayed rules, however, requires recourse to the rule table.

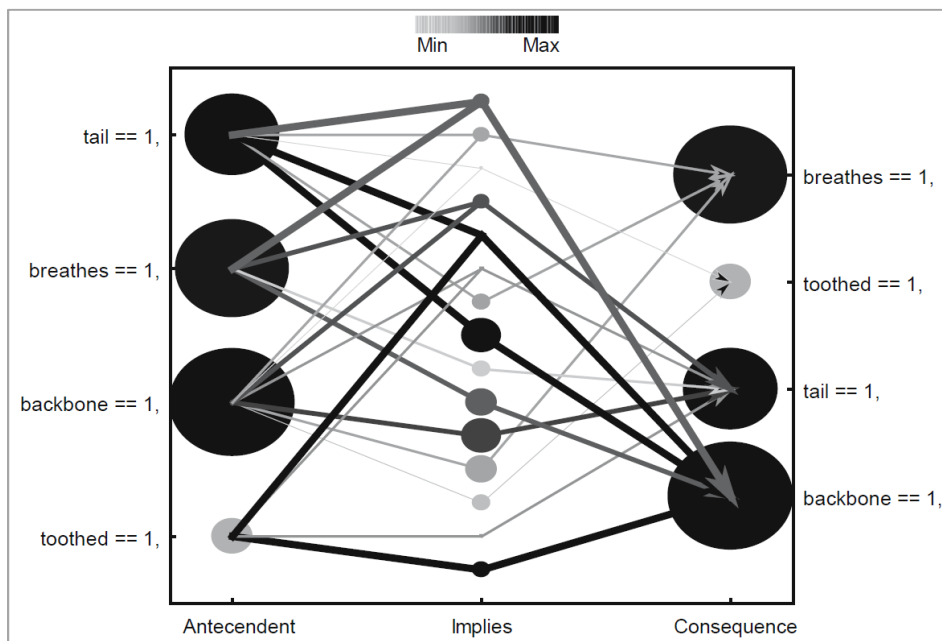


Figure 10: Association Rule Network

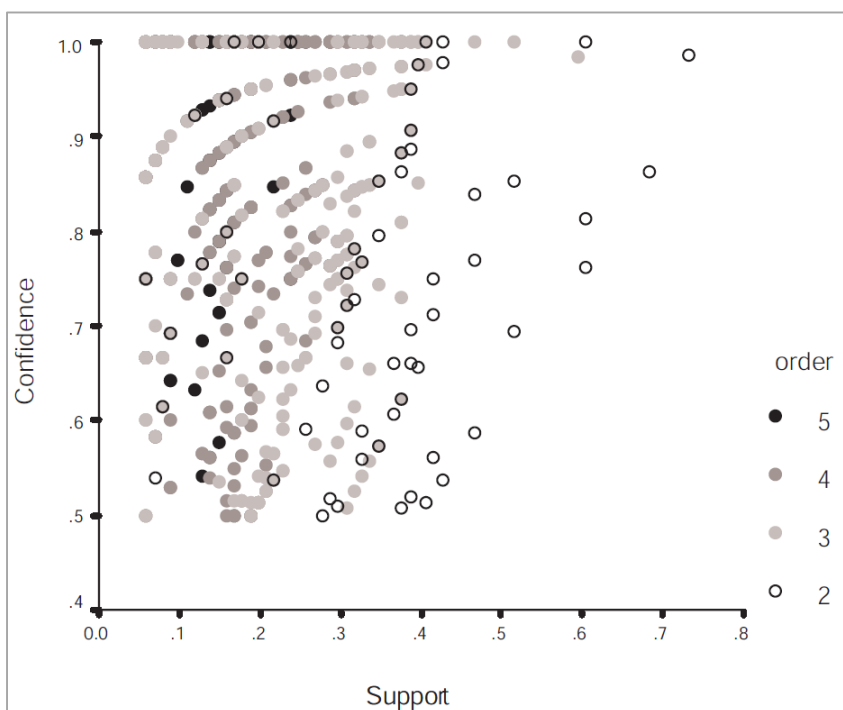


Figure 11: The TwoKey Plot

Another technique for visualizing association rule uses Mosaic plots, or Double-Decker, to provide visualization for single rule and its related rules (Hofmann et al. 2000a; Hofmann et al. 2000b; Hofmann and Wilhelm 2001). With a mosaic plot, all the attributes involved in a rule are visualized by drawing a bar chart for the consequence item and using linking highlighting for the antecedent items. An example is shown in Figure 12. The main drawback of the Double Decker plot is that it is limited to represent a single rule at a time.

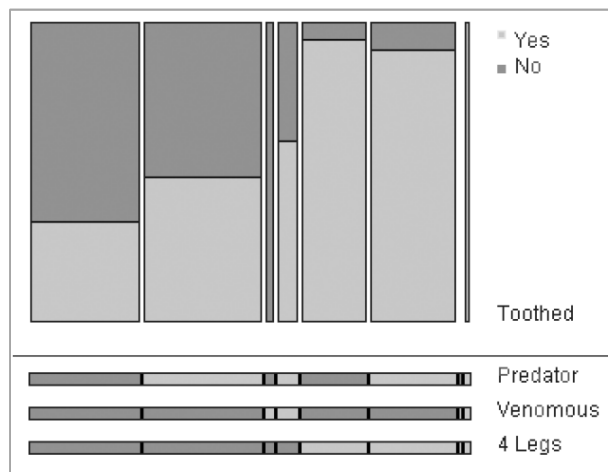


Figure 12: Double-Decker Plot

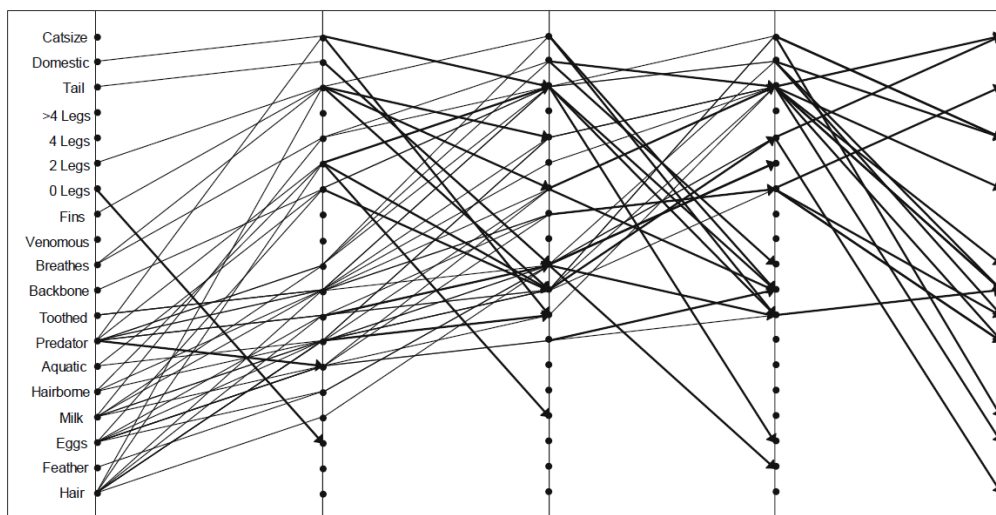


Figure 13: Parallel coordinate plot for association rule visualization

Parallel coordinate plots are another powerful approach used to visualize association rules. The approach was first proposed by (Yang 2008) with the arrangement of items by groups on a number of parallel axes equal to the maximum order of the rules. An example of this type of visualization is shown in Figure 13. A rule is represented as a polyline joining the items in the antecedent followed by an arrow connecting another polyline for the items in the consequence. The items arrangement on each axis should ensure that polylines of itemsets of different groups never intersect with each other. The approach has been further improved and used by several authors (Bruzzese et al. 2003; Bruzzese and Davino 2008; Yang 2008).

Finally, visualizing association rules using factorial method is very promising, especially when interacting with a significant number of rules because it allows to synthesize the information stored in the rules and to visualize the association structure on 2-dimensional graphs. This approach stores synthesized rules in a data matrix where the number of rows is equal to the number of rules and the number of column corresponds to the total number of different items in both the antecedent part (P_{if}) and in the consequent part (P_{then}) of the rules. Also two added columns are used to store confidence and support values. The Multiple Correspondence Analysis (MCA) (Greenacre 1993) is then utilized to analyze this data matrix. Fundamentally, MCA allows a dimensionality reduction on the original variables by identifying the linear combinations of them, the so-called factors. Rules and items are then presented on the reduced dimension subspaces, i.e. the factorial planes. Different views on the set of rules can be obtained by exploiting the results of the MCA process, including item visualization, rules visualization, and conjoint visualization. Items visualization represents the antecedent and the consequent items using the factor

plane where the item points have a dimension proportional to the supports and the confidence. The supports are represented by oriented segments linking the origin of the axes to their projection on the plane, i.e. the plane coordinates. Figure 14 shows an example of item visualization. First, it is easy to identify regions characterized by strong rules. In addition, the closeness between antecedent items and consequent items highlights the presence of a set of rules with a common dependence structure. Visualization for rules can also be performed on the factorial plane, and an example of that is shown in Figure 15. The rules are represented by points with a dimension proportional to their confidence. The proximity among rules indicates evidence of a common structure of antecedent items associated to different consequences. Further examination of the set of selected rules can be carried by using a tabular format. The conjoint visualization of the items and rules is also feasible using factorial planes, as shown in Figure 16. In the conjoint representation, aside from a scale factor, each rule is surrounded by the antecedent items it holds and vice versa each item is surrounded by the rules sharing it. By linking two or more active items, it is possible to highlight all the rules that contain at least one of the selected items in the antecedent.

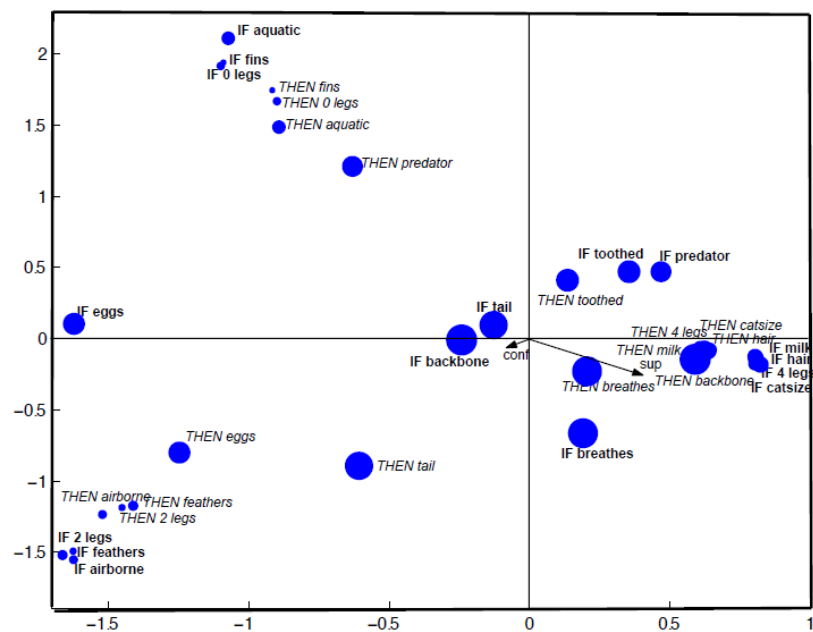


Figure 14: Items visualization

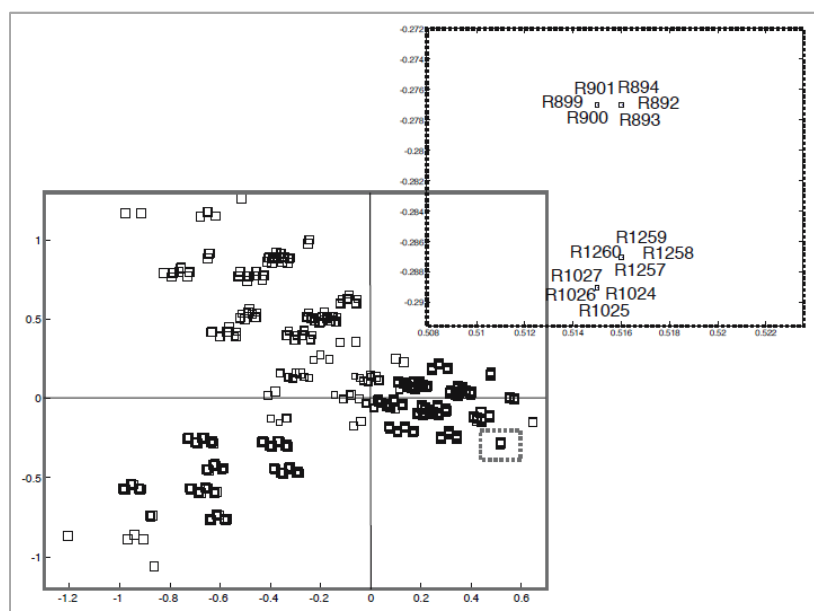


Figure 15: Rules Visualization

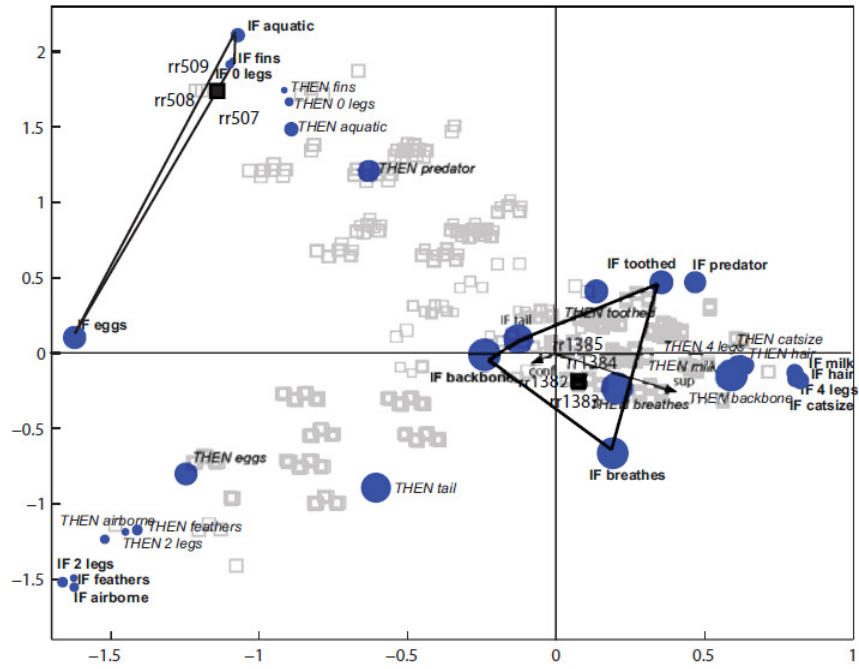


Figure 16: Conjoint Visualization

Although a significant number of visualization techniques have been developed for AR visualization as described above, the majority is not suited to cases with a large number of rules. Moreover, the innovation of these approaches for applications on spatial association rules (SAR) is very much open for discussion. These matters will be discussed within the proposed framework of this research.

2.3.4.2 SAR Evaluation

Once association patterns are mined, their evaluation is indispensable in order to achieve the ultimate goal of the association discovery process, which is to find novel, interesting, and useful association patterns applicable in the geospatial domain. To achieve effective evaluation, human-centered interactivity with visual representations is advised (e.g. Simoff et al. 2008a). Human interactivity often relies on the perceptual, understanding, and reasoning capability of the analyst, which however, largely varies

from person to person. It is, thus, important to establish a set of well-accepted evaluating criteria, or interestingness measure, toward the mined results. Before going into a detailed discussion on rule evaluation based on interesting measures, it is important to mention redundant rules and their elimination.

Redundant rules are rules that are similar to each other and that unnecessarily enlarge the rule set, thus making rule management difficult and cumbersome. Researchers have suggested some solutions for this problem. For example, Cristofor and Simovici (2002) proposed inference rules or inference systems to prune redundant rules and thus present smaller, and usually more understandable sets of association rules to the user. Ashrafi et al. (2004); Ashrafi et al. (2005) presented several methods to eliminate redundant rules. Some of them identify the rules that have similar meaning and then eliminate them. Jaroszewicz and Simovici (2002) presented another solution to the problem using the Maximum Entropy approach with closed form solutions for the most frequent cases.

Regarding interesting measures, the literature suggests two types of measures for association rules, namely objective and subjective (Tan et al. 2006). Objective measures are based on data-driven approach and established using statistical arguments derived from the data. With this type of criteria, the item counts are used; and patterns involving a set of mutually independent items or covering very few transactions are considered uninteresting and eliminated. Because it is domain-independent and requires minimal input from the users, other than to specify a threshold for filtering low-quality patterns, objective measures are easy to derive. Examples of objective interestingness measures include the most popular support and confidence framework. Modifications based on these measures have also been developed. For example, Brin et al. (1997a) identified

correlations and consider both the absence and presence of items as a basis for generating the rules. Chi-squared test for correlation is used to measure the significant of rules. (Omicinski 2003) proposed measures called all-confidence, and bond which are indicators of the degree to which items in an association are related to each other. With all confidence, an association is interesting if all rules that can be produced from that association have a confidence greater than or equal to a minimum all-confidence value. In addition, bond is similar to support but with respect to a subset of the data derived from user-defined conditions rather than the entire data set.

Most objective measures suffer from the rare item problem. A rare set of items is interesting but infrequent. Searching for these rules requires a low support. However, using a low support will generate a large number of non-interesting rules. On the other hand, using a high support reduces the number of rules mined but will eliminate rare rules. Liu et al. (1999) suggested a solution to this problem by allowing users to specify different minimum supports for the various items in their mining algorithm.

Subjective criteria, on the other hand, are established through subjective arguments constructed from domain knowledge including well-accepted theories or conceptual hierarchies. According to subjective criteria, a pattern is considered subjectively uninteresting unless it reveals unexpected information about the data or provides useful knowledge that can lead to profitable actions (Tan et al. 2006). Generally, incorporating subjective knowledge into pattern evaluation is a non-trivial task because it requires a considerable amount of prior information from the domain experts. Existing efforts suggest some approaches to achieve this. For example (Baralis and Psaila 1997) suggested the use of template language to specify a predefined format for different rule

extraction conditions. This means that the association rule templates provide a simplified interface (linkable to visualization) for defining rule extraction criteria. Subjective interesting measures for example based on domain information and knowledge in form of hierarchies are suggested in (Tan et al. 2006). These measures are then used to filter patterns that are obvious and non-actionable.

Evaluation of spatial association rules (SAR) is very similar to that of aspatial association rules (AR) using objective evaluation approaches such as a support-confidence framework. However, like any data mining tasks which involve specific domains of knowledge, SAR evaluation will benefit from subjective evaluation using spatial domain knowledge in order to extract meaningful and interesting patterns. There is also a need to develop frameworks specialized in procedures to couple evaluation approaches with efficient visualization techniques for analytical purposes. Practical solutions to these concerns is one of the focuses of this research and will be discussed in more detail later.

CHAPTER 3: FUNDAMENTALS OF SPATIAL CRIME ANALYSIS

One objective of this study is to implement and test the proposed SpatialARMED framework particularly for use in criminology as a mean of validation. Pertinently, this chapter establishes arguments supporting the development of such framework in relation to the current state of research methodologies in spatial crime analysis and modeling. In addition, the contemporary state of knowledge in spatial crime analysis is reviewed for the purpose of establishing the foundation for SAR result evaluation later on.

3.1 Techniques for Spatial Crime Analysis

Criminal activities in the form of thefts, robberies, assaults, homicides, etc. occur every day almost anywhere in our world and put a strain on the communities, towns and cities in which we live. There are significant monetary costs associated with policing crime and prosecuting offenders. There are also non-monetary social costs associated with crime, which is reflected in changing perceptions of quality of life, mental health and physical security in our daily activities (Murray et al. 2001). Crime analysis thus has attracted the attention of regulators, policy makers, urban planners, and researchers as a vital issue to promote healthy development of any city in the world.

The spatial components in crime activities have been well recognized among criminologists and spatial analysts. This is reflected in the core dimensions of crime, which has been defined as any action against the law (Brantingham and Brantingham (1981)). Crime has four dimensions: (1) a legal dimension (i.e. a law must be broken); (2)

a victim dimension (i.e. someone or something has to be target); (3) an offender dimension (i.e. someone has to do the crime; and (4) a spatial dimension (i.e. the crime has to happen somewhere). Among these, the spatial dimension of crime is regarded as playing a crucial role in understanding crime and how crime can be tackled (Chainey and Ratcliffe 2005). When crime occurs, it happens at a certain location. The offender also must come from a place and this place could be the same or closely connected to the location where the crime was committed (Brantingham and Brantingham 1981; Rossmo 1995). In addition, it is well-known that criminal acts do not occur at random places or random times; instead they tend to occur in certain “zones” of the cities (Burgess 1925; Shaw and McKay 1942). Questions naturally arise as to why it is easier for that person to become a criminal and why that victim at that particular location becomes a target. Looking for the answers to these questions has led researchers on many different pathways of explanations, including factors ranging from internal (i.e. biological and psychological in nature) to external aspects (due to e.g. poor social controls, breakdown of morality in society, feminism-related power struggles, and root causes such as poverty and inequality).

From a spatial modeling perspective, various approaches have been established over the years to explore the relationships between crime and environmental or socio-economic characteristics (Chainey and Ratcliffe 2005). These include exploratory spatial data analysis (i.e. hotspot analysis, spatial dependence estimation) and confirmatory spatial statistical modeling (i.e. spatial regression, geographically weighted regression). Applications of geographical information systems and science for crime mapping and analysis are also well known. As in many other fields, spatial processes such as spatial

dependence and spatial heterogeneity are fundamental to crime pattern analysis (Townsend 2009). Spatial dependence refers to the phenomena in which the crime level in an area is influenced by or at least related to the surrounding area. This often reflects common underlying causes of crime which drive the crime rate or sometimes simply the spillover of crime itself in the neighboring areas. Spatial heterogeneity, on the other hand, refers to the variation of crime concentration across the study area. Spatial dependence and heterogeneity are the main reasons to study crime patterns, in order to discover why some places are victimized more than others, what the associative factors to high or low crime rate are, and whether or up to what degree these factors are spatially dependent or auto-correlated.

In order to identify the concentration of crime, hot spot analysis is often considered. A crime hot spot is defined as a small area with an identifiable boundary containing a concentration of criminal incidents relative to the distribution of crime across the whole region of interest (Anselin et al. 2000). Hot spots permit the rapid identification of the geographic location of crime concentration but they by themselves contribute little to understanding why crime is concentrated in certain locations. However, a visually appealing map can significantly help to identify areas that persistently suffer from crime, and enable a more focused approach to understand areas that require crime reduction resources. This indeed can offer direction for initiating the next analytical stages that explain the problem and how it can be tackled. Hotspot mapping of crime is thus often regarded as the first step toward exploring crime patterns in more detail (Chainey and Ratcliffe 2005). There exist various methods as well as significant efforts in identifying crime hot spots. Point pattern analysis techniques such as quadrat analysis, nearest

neighbour index analysis, and distance-based K-function or areal pattern analysis techniques such as Moran's I, Geary's C and G statistics can be used to test against the existence of spatial dependences and clusters (Jefferis 1999). Continuous surface smoothing techniques, such as kernel density estimation, which use interpolation techniques and include an inverse distance weighting and kriging can be used to visualize the distribution of crime and identify hotspots for the whole studying area (Ratcliffe and McCullagh 1999; Williamson et al. 1999; Chainey et al. 2002; Eck et al. 2005). One can also use more advanced techniques such as Local Indicators of Spatial Association (LISA) statistics (Anselin 1995; Ord and Getis 1995) and the Geographical Analysis Machine (GAM) (Openshaw et al. 1987) to obtain more robustness.

In addition to mapping the concentration of crime, analysts have tried to identify the drivers that potentially contribute to crime. This in many ways helps to determine possible leverage points so that by influencing an underlying driver it may be possible to reduce crime. Features of the physical world (e.g. crime attractors and generators) and socio-economic world (e.g. unemployment, age, heterogeneity, housing tenure, and education) as suggested by the spatial crime theories have been hypothesized to have influence on the incidence of crime. Inferential analysis approaches typically carried out by means of multivariate regression modeling have thus been widespread to test against these hypotheses (LeBeau 1987; Sampson and Groves 1989; Land et al. 1990; LeBeau 1992; Kposowa and Breault 1993; Copes 1999; Rengert and Wasilchick 2000; Potchak et al. 2002). Recently, the role of space for crime analysis has been recognized as central in a number of respects. This in turn has prompted a search for spatial mechanisms such as proximity and diffusion to explain these phenomena (Tolnay et al. 1996; Morenoff and

Sampson 1997; Sampson et al. 1999; Townsley 2009; Zhang et al. 2012). Specialized methods of regression analysis, spatial regression models (e.g. spatial lag model and spatial error model) (Anselin 1988; Anselin 2003) and geographically weighted regression (GWR) model (Brunsdon et al. 1996; Fotheringham et al. 2002; LeSage 2004), have also been utilized for crime analysis to deal with the presence of spatial effects and avoid potentially biased results and faulty inference (Bernasco and Elffers 2010). For instance, Andresen (2006) used a spatial error model to control for residual autocorrelation when analyzing calls for services made to Vancouver police. Deane et al. (2008) also applied this model to a study of city-level robbery rates in 1,056 cities in the United States with 25,000 or more residents. This study, in addition, utilizes various alternative spatial dependence structures based on distance between the cities, as well as state inclusion (i.e., all pairs of cities within a state have a value of 1, and all other pairs have a value of 0 in the corresponding W matrix). Some examples of using spatial lag model are (Baller et al. 2001) for county-level homicide analysis, and (Morenoff et al. 2001; Kubrin 2003) for neighborhood-level homicide analysis. Cahill and Mulligan (2007) applied GWR modelling to analyze violent crime in Portland, Oregon at the block group level, while Malczewski and Poetz (2005) used this modelling approach for studying the spatial variation of the relation between socioeconomic neighborhood characteristics and the burglary risk in London, Ontario.

Recently, developments in computing and spatial data collection techniques have encouraged crime analysts to turn to more advanced analysis techniques such as agent-based modeling (ABM) to allow for more micro level analysis and simulation, and thus build potentially explanatory models of crime. By allowing researchers to create virtual

worlds and inhabit them with simulated populations of heterogeneous, autonomous actors, ABM provides a platform to examine the aggregate level impacts of differing individual level behaviors and thus to explore how the decisions people make on a day-to-day basis translate into observable phenomena (Epstein and Axtell 1996). In comparison to statistical models, ABM has the advantage to consider factors at various levels, including ones which are more localized (i.e. at the level of the individual, street, and neighborhood) and allows feedback loops (Malleson et al. 2009). Fundamentally, ABMs consist of two key components: a population of agents and a simulated environment in which they are situated (Birks et al. 2012). Each member of the population is represented by an autonomous decision making entity with individual characteristics, preferences, and behaviors, the so-called agents. Agent behaviors are defined by a series of simple condition-action rules which are often inspired by existing theories. ABM is often used to explore the patterns from spatial and temporal interactions of multiple heterogeneous agents. Operationally, ABM simulates the progression of time with discrete increments, referred to as cycles. During each cycle, agents perceive, reason, and act based on circumstances and individual characteristics. By manipulating the initial conditions of the ABM and analyzing data collected about the agents and their actions, researchers gain insights into the likely dynamics of certain societal configurations. Examples of employing ABM in criminology involve street robbery analysis (Groff 2007, 2008), exploration of the ramifications of offender mobility patterns (Brantingham and Tita 2008), characterizing patterns of white-collar crime (Kim and Xiao 2008), projecting the likely impact of crime-prevention interventions and policing deployment strategies (Dray et al. 2008; Bosse et al. 2010). Applications of

ABM to crime analysis is still an evolving subfield. Research efforts along this line of research remain aimed at theory confirmation (Birks et al. 2012).

3.2 SAR Mining and Discovery: An Alternative

The limitations of the above techniques lie in their restricted capability to process high volume and nonhomogeneous datasets. In addition, the specific and simple types of relation models that can be accommodated, such as linear regression, and their confirmatory nature is not very robust to handle the complex nature of criminal activities and to allow discovery of unexpected or surprising information. Spatial data mining and discovery naturally come to be a potential solution in these cases. Looking from the data miners' perspective, the complex nature of crime provides great opportunities to test the validity of existing mining algorithms. Association mining and discovery in particular has been recognized as promising toward this tendency (Phillips and Lee 2006; Phillips 2007; Phillips and Lee 2009; Phillips 2009, 2011). It allows the discovery of correlations between crimes and aspatial or spatial dynamics which is a core components of intelligence-led policing, and so permits a deeper insight into the complex nature of criminal behavior, while offering the various advantages of a data mining approach over traditional spatial statistical approaches as previously discussed.

It is not difficult to find existing attempts at making use of data mining approaches for crime analysis, for example (Chen et al. 2004; Keyvanpour and Ebrahimi 2011; Oatley and Ewart 2011) to name a few. One could also find studies directly related to the application of association rule mining and discovery for crime analysis such as (Lee and Phillips 2008; Sathyaraj and Chandran 2010; Lee and Estivill-Castro 2011). Some of these studies take advantages of the off-the-shelf ArcGIS software package to perform

overlay among classified layers of associated features in order to identify the most frequent sets of attributes contributing to crime as shown in Figure 17 and Figure 18. The recent paper by Wang et al. (2013) proposed a spatial data mining framework to detect crime hotspots through their related variables. Basically, so-called Geospatial Discriminative Patterns which essentially are frequent itemsets of crime related variables are identified for areas having relatively high crime intensity. These patterns are then examined for their similarities. The trajectories on map of similar patterns are then used to assist in hot spot detection through iterative hypothesis tests. The focus, however, was not on improving mining efficiency of frequent patterns from a spatial perspective.

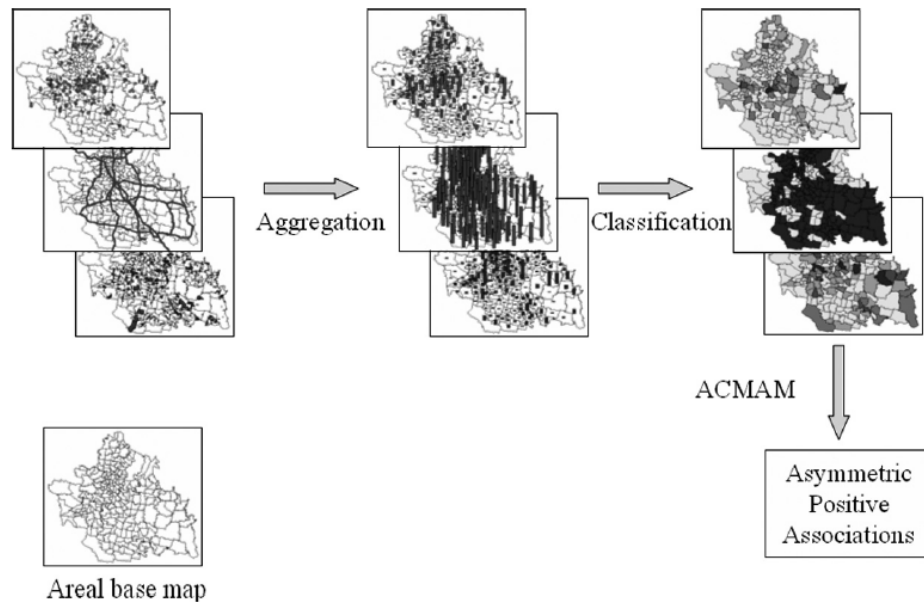


Figure 17: The framework of areal categorized geospatial knowledge discovery (adapted from Lee and Phillips 2008)

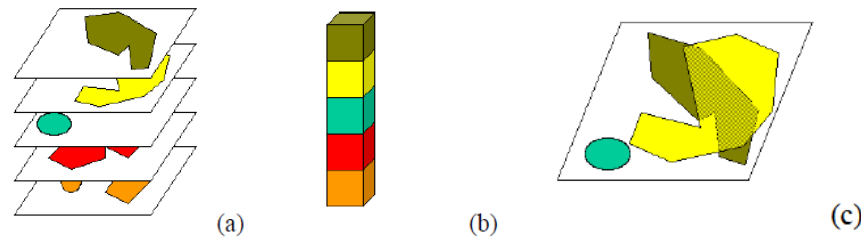


Figure 18: Multivariate association mining technique with (a) ArcGIS feature layers, (b) attribute overlaid for a processing cube, and (c) layer overlaid (adapted from Estivill-Castro 2001; (Estivill-Castro and Lee 2001))

Given the current state of application of association rule mining techniques to crime analysis, it should also be recognized that effective and robust modeling techniques for spatial components of crime and its associates need to be incorporated. In addition, these existing studies lack the capability to validate their findings. This is mainly because evaluation procedures are often ignored. These limitations once again motivate further efforts proposed in this research for enlarging spatial association rule mining and discovery in general and in criminology in particular.

3.3 State of Knowledge in Crime Associations

The current state of knowledge in crime associations is reviewed in this section to establish a foundation for the performance evaluation of the SpatialARMED framework. The review focuses on existing spatial theories and empirical studies indicating factors related to spatial patterns of crime.

3.3.1 Neighborhood Characteristics and Crime

It has been suggested that crime is not equally distributed across the city but localized in certain neighborhoods characterized by economic deprivation, physical deterioration, and social disorders (Burgess 1925; Shaw and McKay 1942). Particularly, communities identified with low economic status, ethnic heterogeneity, residential instability, family

disruption, low level of housing with high population density, low education, unemployment, inequality, and urbanization often experience high level of crime rate. Various cross-sectional empirical studies in fact confirm the existence of these relationships (Richard 1986; Wilson 1987; Sampson and Groves 1989; Patterson 1991; Krivo and Peterson 1996; Morenoff and Sampson 1997; Regoeczi 2003; Lochner and Moretti 2004; Hipp 2007; Walsh and Taylor 2007b; Boggess and Hipp 2010).

One explanation for the link between these neighborhood structural and socioeconomic characteristics and crime lies in the notion of social disorganization. According to Sampson and Groves (1989), social disorganization indicates the inability of a community structure to realize the common values of its residents and maintain effective social controls. A socially disorganized environment is one in which norms and values that support criminal and delinquent behaviour develop due to the lack of effective efficacy or social controls, i.e. the “social cohesion among neighbours combined with their willingness to intervene on behalf of the common good” (Sampson et al. 1997). In the long term, this will support criminal subculture values that remained in the neighborhoods and are passed along to new residents in the process of cultural transmission. According to Sampson, social disorganization is the key to understanding the level of community crime and other social ills (Sampson et al. 1997). From a social disorganization perspective, patterns of delinquency are related to the ecological processes that gave rise to the socioeconomic structure of urban areas and led to concentrated, inner-city poverty and a subsequent breakdown of social orders. The literature consistently shows that community ties with the highest levels of crime and social problems also have the highest rates of poverty (McGhey 1986; Kornhauser 1987;

Brooks-Gunn et al. 1997; Sampson and Raudenbush 2001; Lee et al. 2003 ; Sampson and Raudenbush 2004). Furthermore, some studies suggest that income disparities are linked with high level of crime due to the feelings of relative deprivation, especially when wealthy neighborhoods are located in close proximity to poor ones and poorer individuals may experience feelings of anger and injustice which may ultimately lead to criminal behavior (Morenoff et al. 2001; Hipp 2007). However, the central thesis of social disorganization is that a high rate of delinquency reflects the inability of a community to engage in self-regulation (i.e., social disorganization) and not its economic structure per se. Areas characterized by economic deprivation and physical deterioration tend to have high rates of population turnover, thus residential instability (since they were abandoned by the residents as soon as it was economically feasible) and population heterogeneity (since the rapid changes in composition made it difficult for populations in these areas to make a concerted resistance against the influx of new groups). These factors are highly associated with the destruction of social controls, including as formal as community participations (e.g. attending church together, or local community development program) or as informally as neighborhood watching for each other, and thus are important in relation to crime and delinquency.

Another important explanation is based on the concept of neighborhood effects or peer effects. It suggests that every individual comes with a particular background developed under the strong influence from his or her particular living environment, the so-called neighborhood. From various different angles, the neighborhood is closely identified to an individual as the spatial environment through which the world is perceived during childhood (and thus builds life models, shapes up thinking, and acts

according to the commons); later in life, the neighbourhood also frames daily activities while the individual interacts with peers for the ordinary purposes of making a living, entertaining, and socializing. The neighborhood effects on criminal activities are often found to be particularly strong in communities categorized by a low level of education, poor school attendance, low school quality, high rate of unemployment, and high percentage of single parents (Sampson 1985; Lochner 2004; Lochner and Moretti 2004). Studies in addition show that stronger effects exist in neighborhoods with a high proportion of African-Americans as they traditionally tend to be economically disadvantaged neighborhoods which are also associated with a constellation of other measures of concentrated disadvantages (such as a lower percentage of college graduates, more single parent households, high unemployment, and high rates of poverty) (Adelman et al. 2001; Hipp 2007; Boggess and Hipp 2010).

3.3.2 Routine Activity and Crime

Apart from neighborhood characteristics, the literature also indicates a relationship between daily routine activities of the victims in relation to crimes. Broad changes in contemporary society and the way these changes have impacted on how we live everyday lives may also be highly related to criminal patterns. For example, Brantingham and Brantingham (1984) examined changing patterns of employment and the new criminal opportunities that are created when there are fewer people staying at home during the day. Similarly, Copes (1999) indicated a close link between daily routine activities and motor vehicle thefts.

Explanation for this trend often follows the argument based on routine activity theory which states that the majority of criminals tend to act in a predetermined manner. These

tendencies to react in a similar way to the same opportunities across space are termed aggregated criminal spatial behavior (Brantingham and Brantingham 1984). The idea is that the behavior of victims may significantly explain the occurrence of crime. According to routine activity theory, crime occurs when there is an intersection in time and space of a motivated offender, an attractive target, and a lack of capable guardianship (Cohen and Felson 1979). It should be noted that there is no clear discussion of a target necessarily being a person (i.e. it can be buildings, cars, mailbox, people, or a wide variety of objects and things). There is also no clear definition of guardians. This could mean a person (e.g. police officer, security guard, shopkeeper, or even pedestrian) or CCTV surveillance systems. The routine activity approach is important as time is also considered as significant. People's daily routine activities affect the likelihood they will be an attractive target who encounters an offender in a situation where no effective guardianship is present. Changes in routine activities in society (e.g., more and more women are working) can affect crime rates.

Studies under the umbrella of routine activity theory focus mainly on the nature of targets, of guardians, of offenders, and spatial temporal relationships among them. It is concluded that "Crime opportunity is the least when targets are directly supervised by guardians; offenders, by handlers; and places, by managers" (Felson 1995, p.55). A handler is defined as a person who can influence the behavior of the offender. Such person may therefore be a parent or a teacher for example. A guardian can have some influence over the likelihood of crime. Guardians can be formal, such as police officer, or informal, such as the presence of a friend or pedestrians. The place manager is someone who is able to control a place, such as landlords, street stall owners, store owners, and

ticket clerks. The relationships among these factors can help the analysis of the causes of crime and the mechanisms that can influence those causes. Ultimately, it will contribute to crime prevention.

This theory also indicates that available targets can exhibit different attractiveness to criminals based on the scale of “hot products”. Cohen and Felson (1979) posited that a product is criminally attractive if it is concealable, removable, available, valuable, enjoyable or disposable. Besides, from the spatial perspective, attractive products are not evenly distributed throughout the urban space. While performing everyday routine activity, offenders will search for attractive targets with least guardians and make a decision to commit a crime by weighting up some of the pros and cons. He or she will ask the question of what the rewards are, against the chance of being caught. Saying so suggests that committing a crime is a fairly rational, trying to achieve some sort of desire or goal (Cornish and Clarke 1986; Clarke and Felson 1993).

3.3.3 Environment and Crime

In the literature, the link between environment and crime is well documented. For example, Rengert and Wasilchick (2000) examined the associations to residential burglary behavior and concluded that spatial exploration is very rare in criminal spatial behavior. Most criminals commit crimes in areas with which they are already familiar. (Matthews et al. 2010) suggested that built environment variables were significant predictors of property crime such as residential burglary, non-residential burglary, theft, auto theft, and arson, especially the presence of a highway on auto theft and burglary.

Crime pattern theory (Brantingham and Brantingham 2008) is often found to provide various explanatory mechanisms behind this relationship. By exploring the space

dimension in which crime occurs, particularly the interactions of criminals and their physical and social environments, crime pattern theory helps to examine the “relationship of the offence to the offender’s habitual use of space” (Bottoms and Wiles 2002, p.638).

The first important concept associated with crime pattern theory is cognitive map and awareness space in relation with crime opportunities. Being influenced by the daily activities and routines of their lives, even if offenders are searching for criminal opportunity, they will tend to steer toward areas that are known to them (Brantingham and Brantingham 1984). While offenders perform their daily activities, their repetitive journeys create a “cognitive map” (Brantingham and Brantingham 1984, p.358) of places, routes, and associations. These cognitive maps often contain a list of areas well known to them including both physical infrastructure, such as buildings, travel routes and stops, and social infrastructure, such as a network of connected buddies frequently met at a specific bar. The urban environment that offenders live in becomes a mosaic of places where they have no knowledge intermixed with familiar places. These islands of knowledge and the routes linking them become the “awareness space” (Rengert and Wasilchick 2000, p.61). Crime opportunities are distributed unevenly over space and intersect with an offender’s awareness space; that is where crime happens. There are several reasons why offenders might commit offences in familiar areas. It is helpful for them to know the layout of an area to move around and for a quick get away, if needed. And it has been suggested that offenders often value feeling “comfortable” in an area and not feeling as if they stand out (Rengert 1989; Wright and Decker 1994; Rengert and Wasilchick 2000).

The second crime pattern theory concept is the least effort principle. If there are similar targets in different familiar areas, offenders will often choose the one that requires the least effort to travel to and commit crimes. This is particularly true for instrumental crime, i.e. the crime committed to achieve a goal, and rather less applicable for expressive crime, i.e. more spontaneous, emotional and impulsive crimes done in anger such as violence, rape, assault. By saying so, the distances between offenders and crime sites are usually short. For example, it was found that the average journey from an offender's home to a burglary target was about five kilometers (about three miles) for residential and non-residential burglary (Rossmo 1995; Wiles and Costello 2000). It was also found that many offenders were unemployed and indeed had never worked. They have no resources to venture into unfamiliar areas and their cognitive map was quite small as it did not include a workplace or many recreational opportunities (Wiles and Costello 2000).

The third concept in crime pattern theory is crime generators and attractors. An urban landscape can be perceived as composed of three elements, nodes, pathways, and edges (Lynch 1960). A node is known as an activity place, i.e. a place that an individual is regularly drawn to, such as home, work, or school. As the offender has to travel from one node to another, they use the routes between nodes, which are defined as pathways. Edges exist between different parts of the city. They could not only be physical, such as the boundaries of commercial developments, or the border between a park and a housing complex, but also be perceptual such as borders between areas of different income or racial mix. For offenders, nodes are important as they sometimes tend to be the site of many offences. Pathways provide an opportunity to pass by and scout for new criminal

opportunities. Edges can be barriers to criminal acts in some cases but may also be opportunities in others. For example, if the edge is formed between two income or racial groups, criminals are often unwilling to cross as they would value the comfort of not standing out. However, the case is opposite if there is an expectation that outsiders are usual on the periphery of areas (i.e. the edge) and therefore there is less suspicion against strangers. In terms of crime opportunities, the spatial arrangement of the nodes, pathways, and edges is tied to concepts defined as crime generators and attractors. A crime generator is a particular area or node, where a large number of people is drawn for reasons that are not related to any particular criminal activity that they might commit, but presents conditions conducive to criminal acts such as time and place (Brantingham and Brantingham 1995; Bernasco and Block 2011). Examples of crime generators are shopping malls or parade grounds. Differently, crime attractors are places that create criminal opportunities and attract motivated offenders to the neighborhood or suburb. The lure of a known criminal opportunity draws offenders to the area, enticing them with the knowledge that the area has a reputation for a particular type(s) or illicit opportunity. Examples of these include red light districts, bar districts, and street drug markets.

Crime pattern theory suggests that the level and the type of criminal activity can be generally predicted through an analysis of a city's geographic environment, such as land uses patterns, street networks, and transportation systems. Also, according to this theory, the best way to lower crime is through situational crime prevention in which the focus is not on changing offenders but on reducing the opportunity to commit a crime in a given place.

3.3.4 Thefts of, and from, Motor Vehicles as a Particular Case

Motor vehicle theft (MVT) is widely recognized as a major crime problem in the United States. In 2003, approximately 8.6 billion dollars was lost due to MVT, more than double the 3.3 billion dollars of estimated losses from burglary (Walsh and Taylor 2007b). The public pays both direct and indirect costs. Most of the direct costs are passed on to vehicle owners through high insurance premiums. Indirect costs, although more difficult to estimate, are also significant. They include loss of earnings if the victim misses a day of work, the rental of a temporary vehicle until the car is replaced, the investment in protective devices such as fuel cutoff switches, steering wheel locks, and alarms to protect the replaced vehicle, the investigation of MVT by law enforcement, the prosecution and adjudication of offenders. That is even without talking about the social cost through diminished quality of life experienced by the victim.

In response for the significant impacts of MVT on society, a significant effort has been made in studying MVT, ranging from exploring offender decision-making processes (Copes 2003; Brantingham 2013), understanding and preventing MVT (Maxfield and Clarke 2004), analysing neighborhood structure variables relating to MVT (Copes 1999; Potchak et al. 2002; Rice and Smith 2002; Walsh and Taylor 2007b; Walsh and Taylor 2007a; Roberts and Block 2013), to analyzing locations of thefts and recoveries (Lu and Thill 2003; Suresh and Tewksbury 2013). For this particular study, only the portion of the literature reporting neighborhood socio-economical structural covariates to high MVT is a focus for review.

The neighborhood structural covariates refer to the fundamental demographic fabric of neighborhoods or communities: social economic status (SES), stability, and racial

heterogeneity. A strong connection of these variates in favor of social disorganization and rational choice theories has been acknowledged for general crime analysis. However, it is interesting that the hypotheses of these theories are rather controversial when MVTs are concerned (Rice and Smith 2002). On one hand, past studies on MVT indicated that areas of lower social economic status (SES) and/or high rate of unemployment tended to have more MVTs (Copes 1999; Miethe and McCorkle 2001). In addition, it is also suggested that increased MVT are associated with residential instability, meaning high population mobility, and more single-parent families (Miethe and McCorkle 2001). From a racial heterogeneity perspective, several studies have found vehicle theft rates to be greater in areas characterized by increased racial and ethnic diversity (Sampson and Groves 1989; Clarke and Harris 1992; Bursik and Grasmick 1993; Warner and Pierce 1993). Research by Davison (1995), however, found that MVT was less likely in heterogeneous communities but more likely in predominantly African-American communities. McCaghy et al. (1977) moreover suggested that African-Americans disproportionately commit the offense and a majority of auto thefts are committed by individuals with incomes below the median. The rationality behind these suggestions follow the social disorganization theory which argues that low social-economic status communities and unstable neighborhoods with racial diversity tend to experience informal social break down and have fewer resources to fight against invading criminal elements (Shaw and McKay 1942; Bursik and Grasmick 1993). In contrast, other researches expressed that MVT concentrates among the socially advantaged and thus seems to negate the hypothesis of social disorganization theory. For example, Sanders (1976) states that automobile theft is generally committed by white middle-class youths in groups of two or

more, largely for the excitement. Some others also indicated MVT rates are link positively to their percentage of young male population (Rice and Smith, 2002) and that most MVT occurs for recreational and short-term use rather than for profits. Thus, areas with weaker economies will have lower rates of auto theft (McCaghy et al. 1977; Clarke and Harris 1992).

There also exist studies aiming to understand the spatial autocorrelation of MVT and suggested that factors beyond a community's boundaries may influence that community's crime rate as potential offenders may have search spaces or routine travel patterns spanning several neighborhoods (e.g. Rice and Smith 2002). Walsh and Taylor (2007a, b) have used spatial regression analysis to control for MVT spatial autocorrelation.

In contrast to MVT which has been extensively studied, theft from motor vehicle (TFM) is way far more common but has been taken less seriously in the criminal research community. Apart from a couple of police guide books on thefts of and from motor vehicles (Keister 2007; Clarke 2010), Clarke and Goldstein (2003) particularly analyses TFM for parking lots in the center city of Charlotte, NC but rather focus on the impacts of guardian factors such as fencing, lighting and attendants rather than spatial pattern on TFM. To the knowledge of the author, semantics and spatial associations to TFM remain to be studied in the literature.

CHAPTER 4: RESEARCH QUESTIONS

4.1 A General Framework for SAR Discovery

Chapter 2 of this dissertation not only underscored the potential for association rule discovery to contribute methodologically in a meaningful way in geospatial analysis but also discussed several unresolved issues with this approach, especially in dealing with distinctive characteristics of spatial data and integrating domain knowledge in the discovery process. Existing efforts seem to have overlooked the spatial aspects embedded within the geospatial problems at hand. A major contribution of this research is therefore to address these issues and to propose a comprehensive framework for spatial association rule mining and discovery, dubbed hereafter the SpatialARMED framework.

The development of this framework contributes substantially to the theoretical body of the literature in both spatial data mining and entity-based spatial analysis by facilitating analysis and mining procedures to identify spatial and aspatial associative factors for the phenomenon under study. The framework is seen to be potentially applicable to a variety of domains of application, ranging from social, economic, and public health to seismological and environmental studies, including social analysis, criminology, traffic accident analysis, education performance, health risk analysis, retail marketing, facility management, and ecology. Problems that particularly benefit from this framework can be specified as follows:

- Problems of spatially discrete point data analysis: This concerns a set of point events which for instance represent locations of crime incidents in a neighborhood, of instances of a certain disease, or of traffic crashes in an area. In this case, each event occurs at a particular location falling within an area characterized by one or more variables. Examples of areas include blocks, block groups, counties, districts, and census sub-divisions associated with various socio-economic variables. Functionality is sought to analyze the pattern of the event locations (i.e. clusters of high or low values, their sizes and shapes), identify sets of associative variables to these clusters, and identify spatial interactions between these clusters and sets of collocated geo-features represented by lines or areas.
- Problems of area or network segment data analysis: This concerns events that have been aggregated to a set of areal units (e.g. blocks, block groups, counties, districts, census sub-divisions, etc.) or of network (e.g. streets, rivers, or sewers, etc.) segments. In this case, there are one or more variables whose values are measured over this set of units. The objectives are to detect sets of categorical variables which are often associated with each other, model the spatial arrangement of these values (i.e. identify clusters of high or low values, their size and shape), and detect spatial associative interactions between analysis units and sets of collocated geo-features represented by points, lines, or areas, as well as correlative relationships between analysis units and identified clusters of variables.

It should be noted that the framework proposed herein is applicable to both spatial point processes and areal data. One could carry out point-to-area or area-to-point

transformations using Thiessen polygons or centroids respectively. Similarly, point-to-network mapping can be used to transform point data to network-based data. The basic aim is to understand associations, both spatial and aspatial, of the spatial arrangement of points or areal values, and maybe use this information in making predictions or formulating preventive policies. At the difference from traditional spatial statistical approaches such as spatial regression analysis, this framework utilizes an association rule mining approach and provides innovative visual analytics with knowledge-based evaluation towards the extracted patterns.

The overarching research goal is to determine how to develop the SpatialARMED framework. The central task in this process is to advance an algorithm for spatial association rule mining and discovery which considers essential characteristics of spatial data, including spatial relations and spatial dependencies while emphasizing visual analytics for evaluation. This includes the capabilities (1) to identify spatial dependence structures, including spatial relations and dependencies, that exist in the data, (2) to efficiently represent these spatial components using predicates, (3) to mine spatial association rules, and (4) to evaluate the interestingness of these rules. The specific research questions numbered equivalently to these tasks are:

- 1.1. What spatial components should be considered in SAR mining and how should they be defined?
- 1.2. How to robustly identify the existence and quantify the structure of spatial dependencies (i.e. clusters of dynamic sizes and shapes) for the geospatial phenomena being studied?

- 1.3. How to model spatial relations among geo-features of database as well as relations between features and identified clusters?
- 2.1. How to identify a set of linguistic expressions to represent the extracted spatial relations and spatial dependencies?
- 2.2. How applicable is a fuzzy-set mapping mechanism to map the quantitatively measured values of spatial relations and dependencies to the linguistic expressions defined above, for the purpose of generating spatial predicates with a certain degree of automation? If so, what are the most suitable membership functions?
- 3.1. Is the Apriori-based algorithm implemented in the existing software packages applicable to the SAR mining process? If not, what would be an alternative solution?
- 4.1. How to utilize effective AR visualization approaches, particularly for a significant number of spatial association rules?
- 4.2. How can a geospatial knowledge base be used to assist the SAR discovery process in term of evaluating the mined results? How to establish knowledge-based criteria used in SAR assessment for interestingness?
- 4.3. How can these established criteria be integrated into a visual analytic system?

4.2 Framework Validation

Another fundamental goal is to validate the proposed framework. In this study, the literature reviewed in Chapter 3 points to encouraging prospective frameworks for spatial association rule mining for use in crime analysis, as an alternative to traditional methodologies such as spatial statistical analysis. Thus, the robustness of the proposed SpatialARMED framework particularly for criminology is put to the test so as to endorse

existing spatial crime theories and to discover new crime patterns applicable to the case of Charlotte, North Carolina. The specific research questions are:

- 1.1. Is there any statistical evidence for the existence of high or low crime neighbourhoods in Charlotte, NC? If yes, how to identify the location, size and shape of these neighborhoods?
- 1.2. Can aspatial and spatial associative factors of crime be identified using the SpatialARMED framework? Are these factors consistent with current state of knowledge in spatial crime analysis?

CHAPTER 5: SPATIALARMED FRAMEWORK DEVELOPMENT

5.1 The SpatialARMED Framework

The proposed SpatialARMED framework illustrated in Figure 19 can be briefly described as a composition of five general levels: spatial data and knowledge sources, spatial data analysis, predication, association rule mining, and visual evaluation of mined result.

Regarding the first, the process of SAR mining and discovery starts at the geographic databases and the geospatial knowledge base depicted as the bottom level of the framework in Figure 19. The knowledge base comprises geo-ontologies, concept hierarchies, and to some extent, well accepted spatial associations and collocations reinforced by well-known theories.

The second level performs spatial analysis and seeks to identify spatial components, particularly in term of spatial dependence and heterogeneity structures, prior to executing spatial joins which enable the extraction of spatial relations and regional linkages entailing these identified structures. As discussed earlier, although the SAR mining literature fully recognizes the issues related to unique characteristics of spatial data, approaches to model and incorporate spatial dependence structures into the SAR mining framework remain rather limited. While the tradition of spatial statistical analysis and modeling is to use spatial weight W matrices to capture the spatial dependence structure

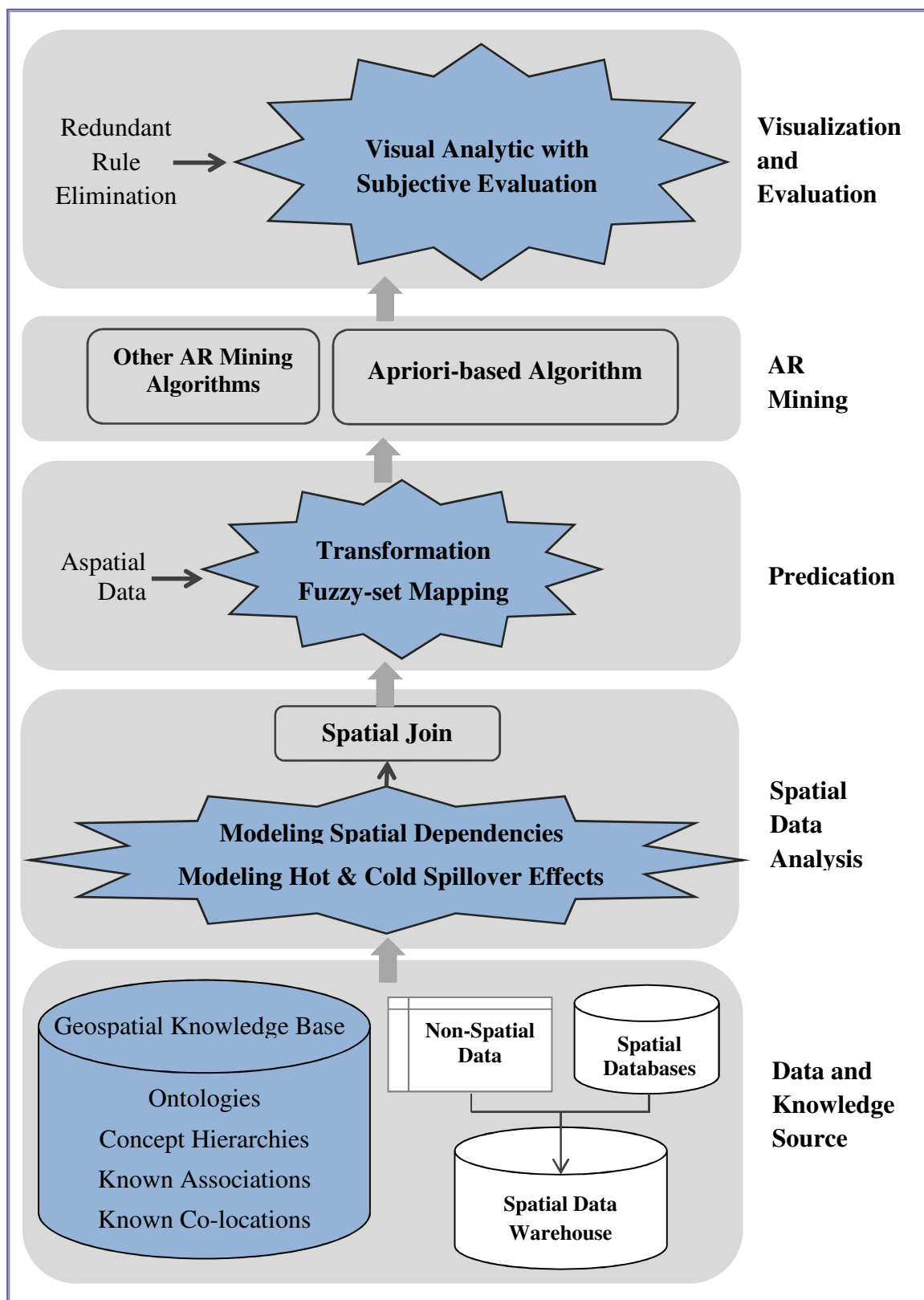


Figure 19: SpatialARMED - a comprehensive framework for SAR mining and discovery

of dependence, SAR mining proceeds differently by not deploying such mathematical expressions but linguistic ones (i.e. spatial predicates). In order to ensure the inclusion of spatial dependence effects, there must be a formal realization on how spatial components are first identified, and then represented within these spatial predicates.

A very similar challenge was encountered by spatial analysts and econometricians over the last three decades, which resulted in what has been referred to as a struggle “with the problem of a proper dependence representation in the W matrix.” (Getis and Aldstadt 2004, p.91). As an essential part of spatial modeling, spatial weight matrices are supposed to be the formal expression (Anselin, 1988), or the theoretical conceptualization (Getis and Aldstadt 2004), of the spatial dependence structures. Apart from very general suggestive rules of thumb are provided for the specification of weight matrices in (Stetzer 1982; Florax and Rey 1995; Griffith 1996), a wide range of visions exist on how to create this theoretical conceptualization. These include making use of spatially contiguous neighbors, inverse distances raised to some power, lengths of shared borders divided by the perimeter, all centroids within distance d , n nearest neighbors, among others. A majority of spatial analysts and econometricians urge special attention to identifying the optimal weight matrices, or risk model misspecification (Stakhovych and Bijmolt 2009; Kostov 2010; Wang et al. 2012). Others have argued that there is “little theoretical basis for this commonly held belief, if estimates and inferences are based on the true partial derivatives for a well-specified spatial regression model” (LeSage and Pace 2010); some even feel it might be best to not use weight matrices at all (Folmer and Oud 2008). For those who are still searching for the correct weight matrix specification, (Getis 2009; Harris et al. 2011) provide excellent reviews.

Back to SAR mining, a similar problem of proper dependence representation is confronted in the spatial predicates. In these cases, the spatial dependence structure ought to be embedded into linguistic-based expressions designed for spatial attributes and spatial relations. For example, considering the rule: if “block-group A is next to block-groups of high crime” then “block-group A has high crime”, the problem of a proper dependence representation is found in the way the predicates “high crime” and “next-to” are defined.

First, how high (or low) should the specifications of “high” (or “low”) are? While in some cases such as human height or ages, common sense could be used to define the break points of these classes, is it more difficult for other socio-economic measures such as income, education, racial heterogeneity, etc... For the later, one might rely on subjective survey-based definitions of high or low, and others argue for the use of more objective based quantization approaches. It is argued here that the definition of high or low should be objectively data-driven derived and tested for significance under the form of clusters, i.e. clusters of high or low values. Thus objects located within clusters of high values will have values classified as “high”, ones being within clusters of low values will have values classified as “low”, and anything lying outside are for medium values.

Second, should the specification of “next-to” be based on the notion of contiguity with a preconceived structure, or distance-related measures, or shared-boundary related measures, or some other specification of greater complexity? Given the nature of spatial predicates and SARs, it perhaps is appropriate to specify the spatial dependence structure as linkages (intra- and inter-) between neighborhood regions (i.e. clusters). The capability to accurately identify the location, size and shape of clusters and to model the spatial

spillover effects of the clusters allows the generation of spatial predicates encompassing characteristics of and spatial relations to cluster-based neighborhoods, thus enabling the discovery of possible functional implications due to spatial spillover effects. In addition, as SAR mining and discovery is often applied to big data, robust, defensible, and data-driven approaches are preferred.

Getis and Aldstadt (2004) suggested an approach of spatial autocorrelation grid-searches using local statistics. By this approach, the spatial autocorrelation structure embedded in the data can be extracted, and subsequently, used for creating spatial weight matrices or for identifying clusters. In 2006, this approach is explicitly demonstrated under the form of an algorithm named AMOEBA (A Multi-directional Optimum Ecotope-Based Algorithm). Rogerson and Kedron (2012) applied this approach to examine the optimal weights for focused tests of clustering and referred to it as desirable. Being not only methodologically sound and robust in identifying irregular spatial clusters (i.e. whatever forms empirically exist in the data), an AMOEBA-based approach is also flexible for use with various data types (i.e. point, polylines, and polygons). Although this approach is potentially computationally expensive, high performance computing can be deployed to gain better performance.

The next level of the framework deals with the formation of predicates (i.e. predication), for both aspatial and spatial components. This involves transformations of the quantitatively measured (e.g. numeric) values of these components to qualitative linguistic expressions (e.g. categories of low and high, near and far, or big and small, etc.). In some cases, geo-ontologies and semantics play an important role during this

process, especially in defining membership functions. The SpatialARMED framework applies principles of fuzzy-set mapping to accomplish this task.

The third level of the framework, data mining, focuses on finding frequent sets and generating rules using the set of aspatial and spatial predicates generated in previous level. The literature suggests various different association rule mining algorithms as reviewed in Sections 2.2 and 2.3 of Chapter 2. The SpatialARMED framework utilizes the existing Apriori-based approach implemented in the LUCS KDD-ARM software package (Brin et al. 1997b; Coenen 2004) along with support and confidence cut off strategy for finding frequent item sets and association rules.

The final level of this framework comprises interactive visual analytics to support the evaluation of mined rules. Although significant efforts have been devoted to association rule visualization as reviewed in Section 2.3.4 of Chapter 2, most of them are not capable of handling a large rule set and to provide a user friendly environment for interaction. Thus, they stop short of providing a true visual analytic environment. Some advanced visualization approaches such as factorial methods (Benzecri 1973; Greenacre 1993) that support sub-group visualization are more effective. Such visual analytic systems, however, require the availability of proficient evaluation criteria for rule interestingness in order to facilitate sub-group division and ensure discovery efficiency. The evaluation criteria basically provide standards to assess the interestingness of a rule; this is often based on statistical measures for the dataset in use (i.e. objective types) or based on existing knowledge (subjective types) as elaborated in Section 2.3.4 of Chapter 2.

While the field of visual analytics or visual data mining has been fast emerging over the past few years, literature shows very limited discussion related to the formalization of

the subjective criteria for the evaluation of spatial association rules. This research aims at enhancing an existing AR visualization approach by proposing a set of subjective criteria for SAR evaluation based on geospatial domain knowledge.

To illustrate the algorithms and specifications associated with each level of the SpatialARMED framework, a common example is used hereafter. This concerns a geospatial dataset containing a set of participating feature classes A, B, and C, part of the so-called dataset ABC. One wishes to discover aspatial and spatial associations to C derivable from this dataset. Therefore, C is regarded as the reference feature, while A and B are task-relevant. In order to implement the SpatialARMED framework, units of mining (UoM) are required to be identified (or constructed). In a relational database, these UoMs are equivalent to the so-called “tuples”. In a single relational database (i.e. table), the UoMs (or tuples) are basically represented by the rows of the table with unique identifications. In a spatial database, the UoMs take the form of either points (e.g. houses), line segments (e.g. street segments), or polygons (e.g. census tracts). Association rules are then mined by regarding these UoMs as transactions and the mapped attributes (aspatial and spatial) as items.

Figure 20 illustrates dataset ABC. In this case, A is a set of all houses within the study area. B is the set of all shopping centers with related information such as their sizes and popularity ranks. C is a set of census blocks with demographic information such as income and employment. A single relational spatial database can be constructed for this case as shown in Table 1 in order to facilitate association mining. The UoMs in this case can be blocks. Using the SpatialARMED framework, the objective is to find aspatial and

spatial association rules involving not only the aspatial attributes of the participating features but also the spatial components that exist within or between them.

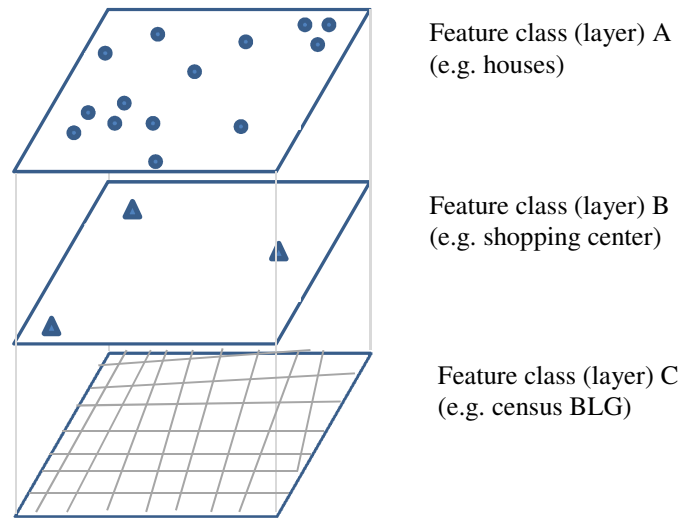


Figure 20: Spatial data set ABC for the SpatialARMED framework demonstration

Table 1: A single relational table derived from dataset ABC

Predicates UoA	A-related				B-related			C-related			
Block 1											
...	⋮	⋮	⋮		⋮	⋮		⋮	⋮		

5.2 Mining Spatial Dependence Structure and Modeling Spillover Effects

The ultimate goal of SAR mining with SpatialARMED is to discover spatial associations, i.e. the functional implications due to spatial relations or spatial spillover

effects among participating factors. In order to assure accurate results, identifying spatial dependence structure involving concentrations of high values (i.e. hot spots) and of low values (i.e. cold spots) as well as modeling the spatial spillover effect of these spots for each and every factor become a critical task for SpatialARMED. From the perspective of predication, the capability to accurately identify the location, size and shape of these spots (i.e. clusters) allows the generation of spatial predicates expressing “within” or “contain” relations. Modeling the spatial spillover effect of the clusters, in addition, allows predicates of “near-by”, “next-to”, “under-strong-impact”, “under-low-impact”, etc. relations. SpatialARMED differs from existing approaches in SAR mining due to its capability to mine whatever spatial dependence structure exists in the data, rather than utilizing predetermined concepts. Moreover, for the first time in the SAR mining literature, spatial spillover impacts are modelled and integrated into the predication process. The following subsections provide detail discussion on algorithms and implementation aspects of these particular tasks.

5.2.1 AMOEBA for Spatial Dependence Structure Quantification

AMOEBA (A Multi-directional Optimum Ecotope-Based Algorithm) is originally proposed in (Getis and Aldstadt 2004; Aldstadt and Getis 2006) to model spatial autocorrelation using local statistic tests. Although designed for areal analysis, AMOEBA deployment on spatial point processes is feasible by point aggregation to meaningful areal or network-segment units.

Functionally, starting from one or more “seed” spatial units, AMOEBA searches and tests for local spatial dependence in all directions until revealing the totality of the spatial dependence that is subsumed in the data. Local standardized Gi^* statistics are used to test

for spatial dependence. A positive G_i^* indicates that there is clustering of high values around analysis unit i ; a negative number indicates low values. The G_i^* values are estimated cumulatively around each “seed” observation as distance increases from it. When these values fail to rise absolutely with distance, the cluster diameter is reached. This implies that any continuity in spatial association or dependence over distance ends at that diameter distance. This distance is often called critical distance, d . For a given location i , the statistic G_i^* is defined as:

$$G_i^* = \frac{\sum_{j=1}^N W_{ij} X_j - \bar{X} \sum_{j=1}^N W_{ij}}{S \sqrt{\frac{N \sum_{j=1}^N W_{ij}^2 - (\sum_{j=1}^N W_{ij})^2}{N-1}}}$$

where N is the number of spatial units, X_j is the value of the phenomenon of interest at location j , \bar{X} is the mean of all the values, W_{ij} is an indicator function that is one if unit j is in the same designated region as unit i and zero otherwise and

$$S = \sqrt{\frac{\sum_{j=1}^N X_j^2}{N} - (\bar{X})^2}.$$

The null hypothesis for a test based on this statistic is that there is no spatial dependence between the value found at a site and its neighbors within the designated region.

The AMOEBA approach based on G_i^* can be described as follows: At the outset of the AMOEBA procedure, the G_i^* value for the spatial unit i itself is computed. This value is denoted $G_i^*(0)$ and the cluster consists of just the i^{th} unit. A $G_i^*(0)$ value greater than zero indicates that the value at location i is larger than the mean of all units and, correspondingly, a value less than zero indicates that the value at location i is smaller than the mean. The next step is to compute the $G_i^*(1)$ value for each region that contains i

and all combinations of its contiguous neighbors (see Figure 21a). The combination that maximizes the absolute statistic $G_i^*(1)$ becomes a new high or low value cluster. At each succeeding step, contiguous units that are not included in the cluster are eliminated from further consideration. Likewise, units included in the cluster remain in the cluster. Subsequent steps evaluate all combinations of contiguous neighbors and new members of the cluster are identified. This process continues for k number of links, with $k = 2, 3, 4, \dots, \max$ (see Figure 21b). The final cluster (k_{\max}) is identified when the addition of any set of contiguous units fails to increase the absolute value of the G_i^* statistic. Figure 21c shows a complete AMOEBA cluster in a raster setting. The maximum number of links in this case is five ($k_{\max} = 5$). After ecotopes for each and every cell within the study area are identified, the AMOEBA algorithm is continued by keeping the non-overlapping ecotopes with the highest G_i^* values. Final ecotopes (or clusters) are reported as the result of performing Monte Carlo-type permutation test to calculate the statistical significance of each ecotope.

The complication of AMOEBA-based spatial dependence structure identification depends very much on the data size, i.e. the number of cells over which ecotopes should be identified, and the configuration of the spatial dependence structure involved, i.e. the number of neighbors whose combinations should be tested for significant G^* . A parallel computational implementation of the AMOEBA algorithm is possible for large databases in order to increase the computational efficiency (Widener et al. 2012).

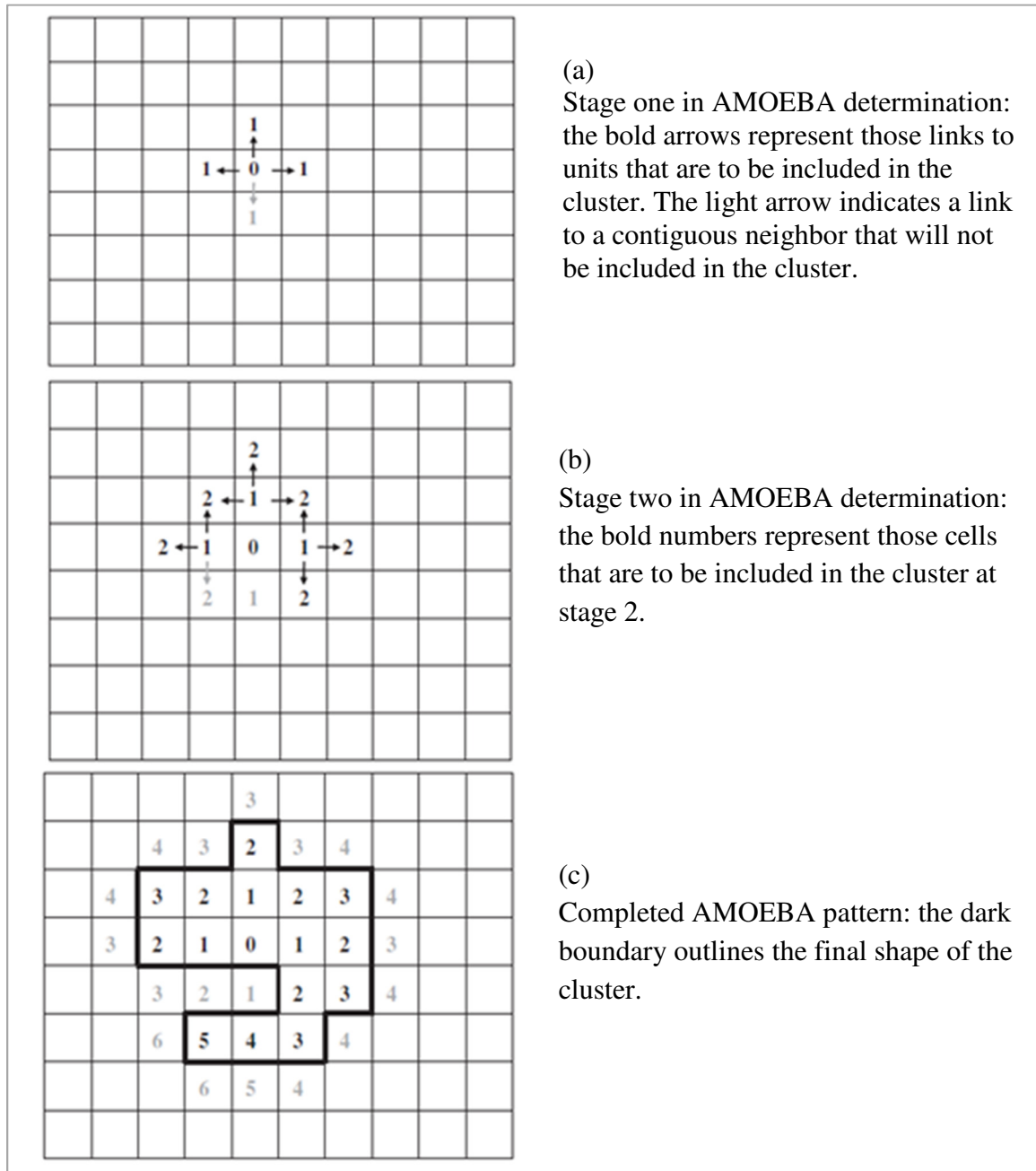


Figure 21: AMOEBA procedure (adapted from Aldstadt and Getis 2006)

Fundamentally, the parallelization of the AMOEBA process involves two phases as shown in Figure 22. The first deals with data decomposition and distribution to parallel high performance computing (HPC) cores while the second recompiles the parallel HPC outputs, performs ecotope overlapping, carries out significant tests, and generates final

spatial dependence structure, i.e. clusters of high and low value concentrations with associated G^* values and cell members.

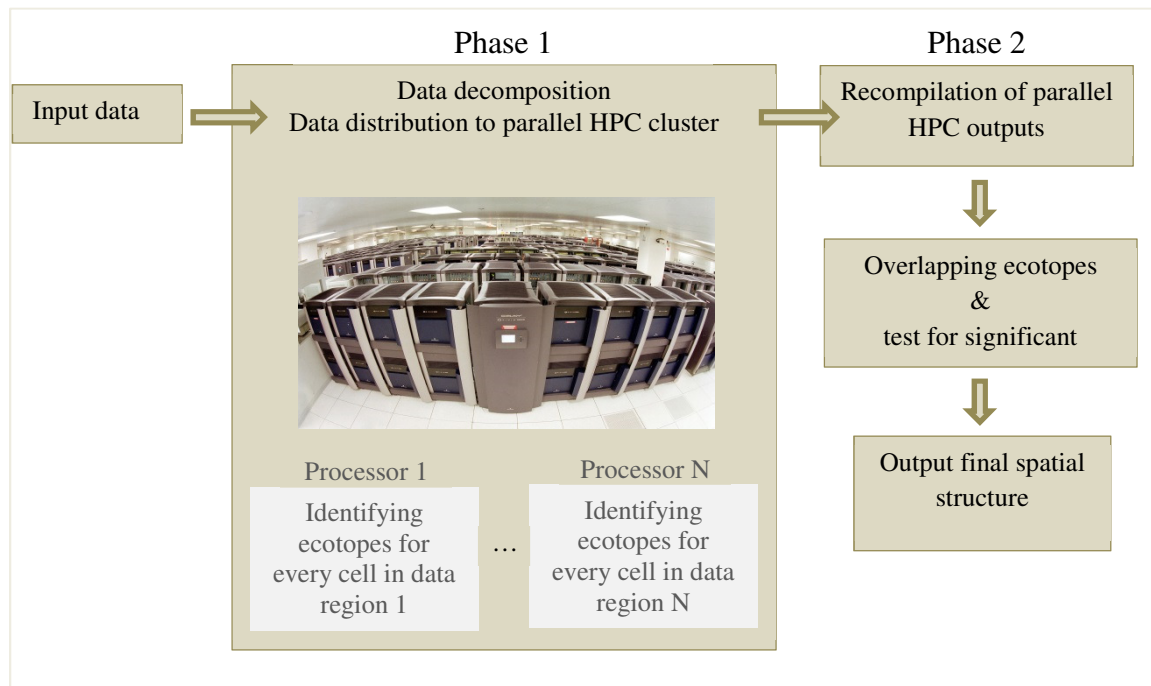


Figure 22: AMOEBa parallel computing workflow

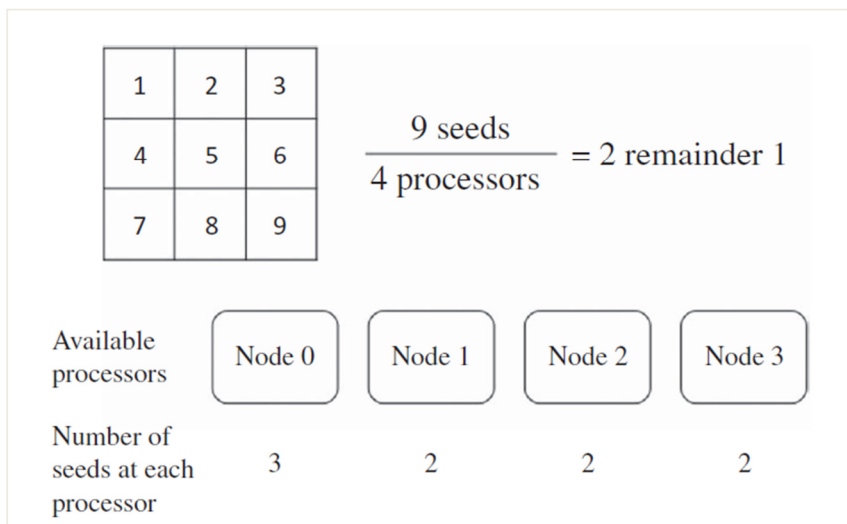


Figure 23: AMOEBa parallel processing phase 1 with regional decomposition

In phase 1, a regional decomposition scheme (Widener et al. 2012) as described in Figure 23 can be chosen to decompose the input data. With this approach, each computing processor is aware of the number of seeds (or cells) it is responsible for. The parallel data decomposition algorithm passes a starting location and end location to each processor, that delineate the region of the dataset it must calculate. The AMOEBA algorithm is set up to run at each processor to identify ecotopes for every seed within the sub-region it is responsible for, and output the results into text files. These text files are input into data recompilation algorithm in Phase 2 which performs ecotope overlapping and significance testing.

Performing spatial clustering on various aspatial attributes allows the identification of neighborhoods characterized by specific factors such as income (i.e. rich or poor), race (i.e. black, white, or Hispanic), house tenure (i.e. own or rent) or crime (i.e. high crime or low crime), to name a few. Referring to the example of using dataset ABC, spatial analysis using AMOEBA identifies clusters of houses with high or low values as well as clusters of BLG-based population with high or low incomes as shown in Figure 24. Accurate information on the cluster sizes is also obtained as a result of the AMOEBA procedure.

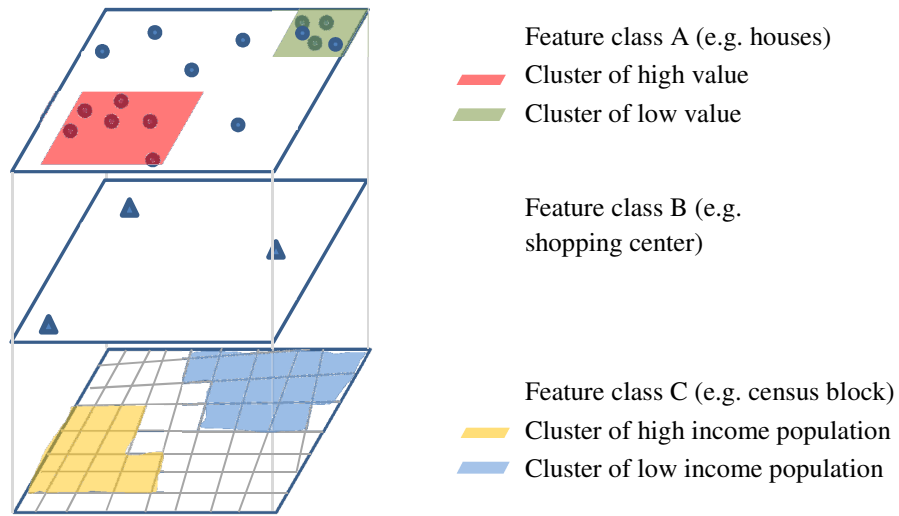


Figure 24: An example to demonstrate the AMOEBA-based clustering result for the dataset ABC

5.2.2 Spatial Dependence Structure Spillover Effect Model

The subsequent step in the SpatialARMED framework is to model the spatial spillover effect of the identified clusters. While the purpose of performing spatial clustering is to identify the exact location and dimension of the clusters, the objective here is to quantify the spatial diffusion impact of these clusters. Integrating these impacts in SAR mining entails the capture of functional associations due to indirect spatial impact according to the Waldo Tobler's first law of geography, "everything is related to everything else, but near things are more related than distant things." For instance, with the ABC dataset, one will certainly be interested in rules related to house units located within the high income neighborhood but also, in the ones related to houses located next-to or under the influence of high income neighborhoods. So the "next-to" or the influence impact should be properly modelled for accurate implication.

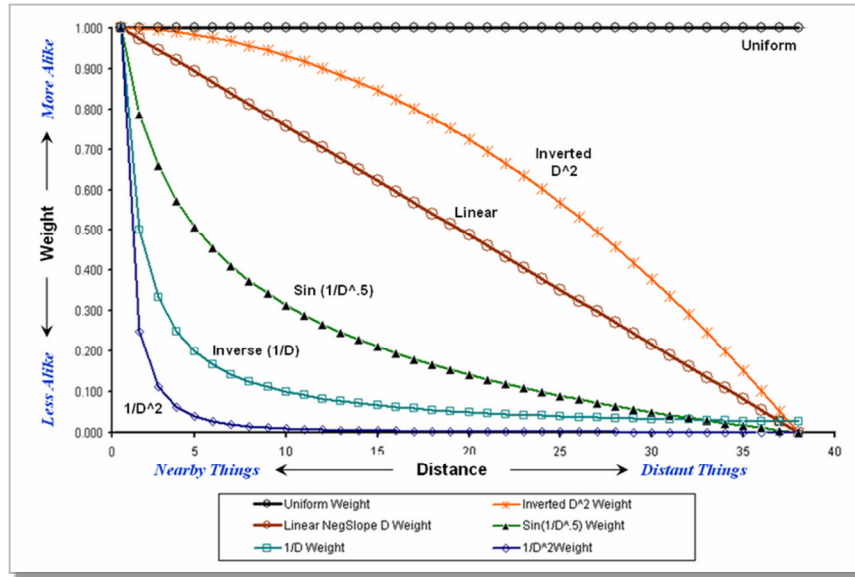


Figure 25: Distance weighted spillover effect models

The challenge of this task involves the identification of a model which accurately represents the spillover impact of the clusters over space, which substantially depends on the size, shape, and internal structure of the clusters as well as the nature of the spillover phenomenon itself. Distance decay functions implying spatial diffusion at different rates, as shown in Figure 25, should be taken into consideration during model construction. The way we model the rate of change in the spatial spillover effect regarding to a particular source of diffusion has a significant impact on the values of the spatial predicates, for instance, “next-to” or “under-influence-of” this source. A sensitivity analysis is recommended then in order to examine the impacts on mined associations of alternative specifications of the spillover effects such as different distance-based diffusion rates.

As the result of the AMOEBA-based spatial clustering process, each of the clusters is associated with a list of members, the standardized cluster G^* value indicating the level of concentration, or high or low values. Consider an example dataset with variable V for which the AMOEBA spatial clustering algorithm identifies a set of positive G^* clusters,

$C_P = \{Hv_1, \dots, Hv_N\}$. Each of these clusters Hv contains m members, $Hv = \{i_1, \dots, i_m\}$. As the simplest case, the spatial spillover effect of concentrations of high values on variable v at location j , $sim_P(j)$, can be modelled as a function expressing the diffusion of cluster G^* values over space, i.e. cluster G^* divided by distance to the cluster centroid raised to power α as shown in Equation (1) and Figure 26.

$$sim_P(j) = \sum_{i \in C_P} \frac{G_i^*}{d_{ij}^\alpha} \quad (1)$$

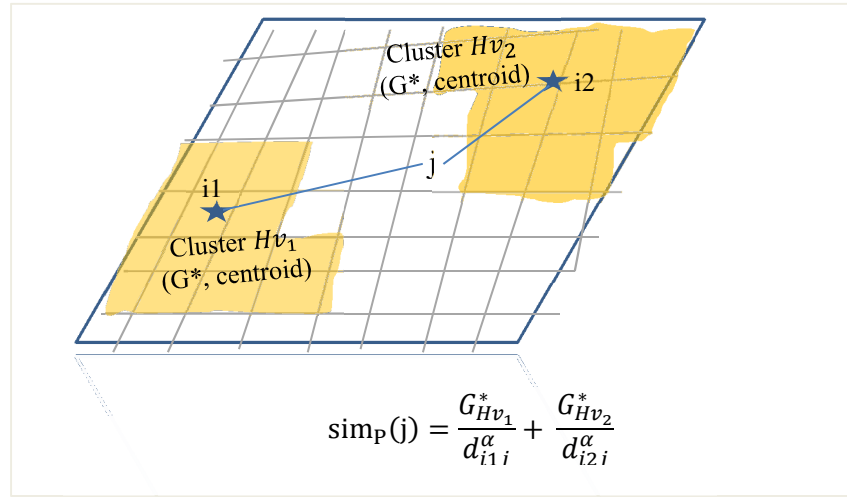


Figure 26: Modeling the spatial spillover impact using cluster G^* values

A similar formulation can be used to model the impact of negative G^* clusters, $C_N = \{Lv_1, \dots, Lv_s\}$. Each of these clusters Lv contains k members, $Hv = \{i_1, \dots, i_k\}$. The spatial spillover effect of concentrations of low values for variable v at location j , $sim_N(j)$,

$$sim_N(j) = \sum_{i \in C_N} \frac{G_i^*}{d_{ij}^\alpha} \quad (2)$$

The drawback of this model is that it overlooks the spatial dimension and internal structure of each cluster. Although the cluster G^* value indicates how strong the

concentration of high or low values is, considering the cluster as the whole, it does not inform about the homogeneity of high or low values contributed by cluster members (i.e. information regarding high or low value distribution within the cluster itself). In addition, using cluster centroids for spatial diffusion modeling leaves out impact differences due to cluster sizes and shapes.

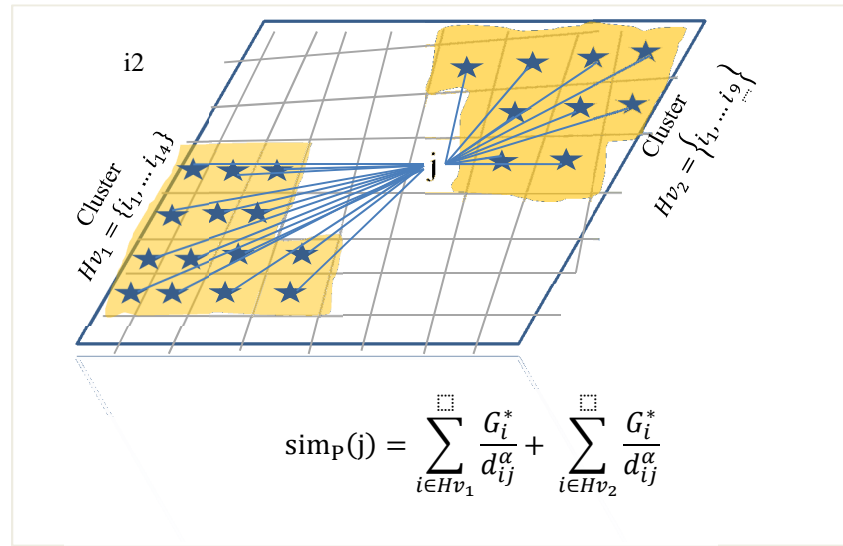


Figure 27: Modeling the spatial spillover impact using individual G_i^* values

An alternative is to consider every member (i.e. cells) of each cluster and use their individual G^* values instead of the cluster G^* values to model the spatial spillover impacts as demonstrated in Figure 27. By this way, cluster size, shape, and internal variation of the individual G^* values within a cluster will be taken into account. Spatial spillover effects of a variable v at location j , $\text{sim}_P(j)$ and $\text{sim}_N(j)$, can then be modelled as:

$$\text{sim}_P(j) = \sum_{Hv \in C_P} \sum_{i \in Hv, i \neq j} \frac{G_i^*}{d_{ij}^\alpha} \quad (3)$$

$$\text{Sim}_N(j) = \sum_{Lv \in C_N} \sum_{i \in Lv, i \neq j} \frac{G_i^*}{d_{ij}^\alpha} \quad (4)$$

where C_p is the set of all clusters of high values in the study area, C_N is the set of clusters of low value in the study area, Hv is a cluster of high value, Lv is a cluster of low value, and i is a cluster member.

Apart from modeling the spatial spillover of hot and cold spots, it is proposed here for the SpatialARMED framework to consider the spatial spillover impact of point facilities, referred here as points of interest (POI) which could potentially have associations to the phenomenon under study. For example, consider the dataset ABC when there is a need to model the spillover impact of shopping mall i to a particular location j , to generate predicates related to the potential population of shoppers. Particular characteristics of the malls could be considered in the model. These could potentially be the size, number of shops it contains, or its popularity index, to name a few. This is useful, for example from the perspective of crime generators or crime attractors. Generally, a gravity-type model can be applied to model the spillover impact of POI as follows:

$$\text{Sim}_{POI}(j) = \sum_{i \in \text{POI}} \frac{A_i}{d_{ij}^\alpha} \quad (5)$$

where POI is a set of POIs, A_i is the attribute of interest, and d is distance.

Due to the complexity of spatial diffusion for different phenomena, examination using different types of models is highly recommended.

5.3 The Process of Predication

Predication is the process of generating predicates expressing aspatial or spatial attributes of both reference and relevant variables for SAR mining. In the SpatialARMED framework, this process is facilitated by the miner's decision on choosing the unit of tuples, along with spatial join operations to extract spatial relations among mining objects

based on the spatial dependence structures and their spillover effects identified from the previous step, and a numeric-to-nominal mapping mechanism for predicate values.

5.3.1 Unit of Tuples and Effect of MAUP in SAR Mining

The very first task in the predication process is to identify the unit of tuples, i.e. the unit of each record in the relational table to mine. For example with the dataset ABC, the unit of tuples could be houses, with house price as decision attribute. So predicates for each house will be generated and recorded as a row in the final table to mine. Alternatively, the unit of tuples could be areal block groups and each row of the final table to mine contains attributes for one block group. In such a case, one could generate a variable expressing the number of expensive houses, e.g. with price more than \$2,000,000, located within each block group.

Data manipulation is important during this process and SAR miners need to be aware of the modifiable areal unit problem (MAUP) (Openshaw 1983). This problem will take its effects in SAR mining when miners use data of different spatial resolutions and chose the one with highest resolution as unit of tuples. In this case, data with lower spatial resolution is mapped onto the one of higher resolution with a one-to-many relationship. This causes the frequent item set count and confidence estimation, and therefore rules, to be biased in SAR mining. In order to avoid this issue, it is recommended to use data with the lowest spatial resolution as unit of tuples when mining spatial association rules.

5.3.2 Spatial Join Operation

The purpose of the spatial join process is to extract all the spatial relations between features associated with reference attributes and those with task-relevant attributes, including spatial dependence structures identified from AMOEBA-based clustering.

Theoretically, several spatial relationships between objects exist, including topology-based, distance-based, and direction-based. The topological relationships are invariant under homeomorphisms, such as rotation, translation and scaling. Their semantics is precisely defined by means of the nine-intersection model proposed by (Egenhofer and Franzosa 1991). The distance between two points is typically computed based on Euclidean measures, while the distance between two geometries (e.g., two areas) is defined by some aggregate functions (e.g., the minimum distance between two points of the areas). Distance relationships can be non-metric, especially when they are defined on the basis of a cost function which is not symmetric (e.g., the drive time). Directional relations can be expressed by the angle formed by two points with respect to the origin of the reference system or by an extension of Allen's interval algebra, which is based on projection lines (Mukerjee and Joe 1990). ArcGIS software and openGIS source codes offer a variety of implemented functions and operations which can be used to perform spatial joins in the process of SAR mining and discovery. Details in terms of techniques and algorithms for these functions and algorithms can be found in (Jacox and Samet 2007). For large geospatial databases, some advanced algorithms can be utilized to enhance the spatial join process in SAR mining, as reviewed in Section 2.3.3.

SpatialARMED, in particular, focuses on the extraction of spatial relations in the context of the AMOEBA-based spatial dependence structures of hot and cold spots as well as with the spillover impacts of these spots. As indicated earlier, this allows the generation of predicates expressing spatial relations, such as “within”, “next-to”, “under-effect-of”, among the mining objects and associated features.

Table 2: An example of spatial join within the SpatialARMED framework

UoA \ Predicates	A-related	B-related	C-related			
Block 1	High % of area overlapped with neighborhood of expensive houses (%)	Located under strong influence of popular mall impact (SIM)	Located within neighborhood of high income (G^*)	Located within neighborhood of low employment (G^*)	Located under strong influence of racial heterogeneity (SIM)	Has high criminal activity (G^*)
...	

Concerning the example of dataset ABC, the spatial join process results in the formation of spatial relationships among features A, B, and C. with respect to the identified “neighborhoods”, i.e. clusters, of expensive houses, of high income, of low employment, of high crime or the spillover effects of popular shopping malls and ethnic heterogeneity, as described in Table 2.

5.3.3 Numeric-to-Nominal Mapping Mechanism

As discussed earlier in Section 2.3.2, the AR literature recognizes that fuzziness may be an issue with association rules because linguistic expressions are used in both aspatial and spatial predicates. This is the result of a process in which transformation from quantitative numeric ranges into qualitative linguistic-based categories are required. For instance, with the example using dataset ABC, in order to mine association rules out of Table 2, one needs to categorize the values of attributes into “EXPENSIVE”, “HIGH”, AND “STRONG INFLUENCE”.

In the SpatialARMED framework, fuzzy-set mapping is used to perform such transformation. This results in the use of so-called fuzzy spatial association rules. A review of the extant literature along this line of research was provided in Section 2.3.2. By definition, a fuzzy set is a set without a crisp, clearly defined boundary. Fuzzy sets are used to describe vague concepts while admitting the possibility of partial memberships to several concepts. The degree an object belongs to a fuzzy set is denoted by a membership value, or degree of membership, between 0 and 1. A membership function (MF) associated with a given fuzzy set maps an input value to its appropriate membership value (Klir et al. 1997). The membership function fundamentally is a curve whose shape is specified to suit the mapping logic while allowing simplicity, convenience, speed, and efficiency. For applications which involve the determination of complex membership functions, adaptive training algorithms can be optimized. For this study in particular, if one considers the standardized distribution of a certain attribute, a fuzzy mapping mechanism using trapezoidal membership function for a three-category classification defined on the range of variation of this attribute as shown in Figure 28 can be used. This approach is widely used and easy to implement; also, this type of function is very efficient in mapping numeric data range into High-Low categories based on fuzzy thresholds which is often the case for socio-demographic data and distance-based diffusion effects.

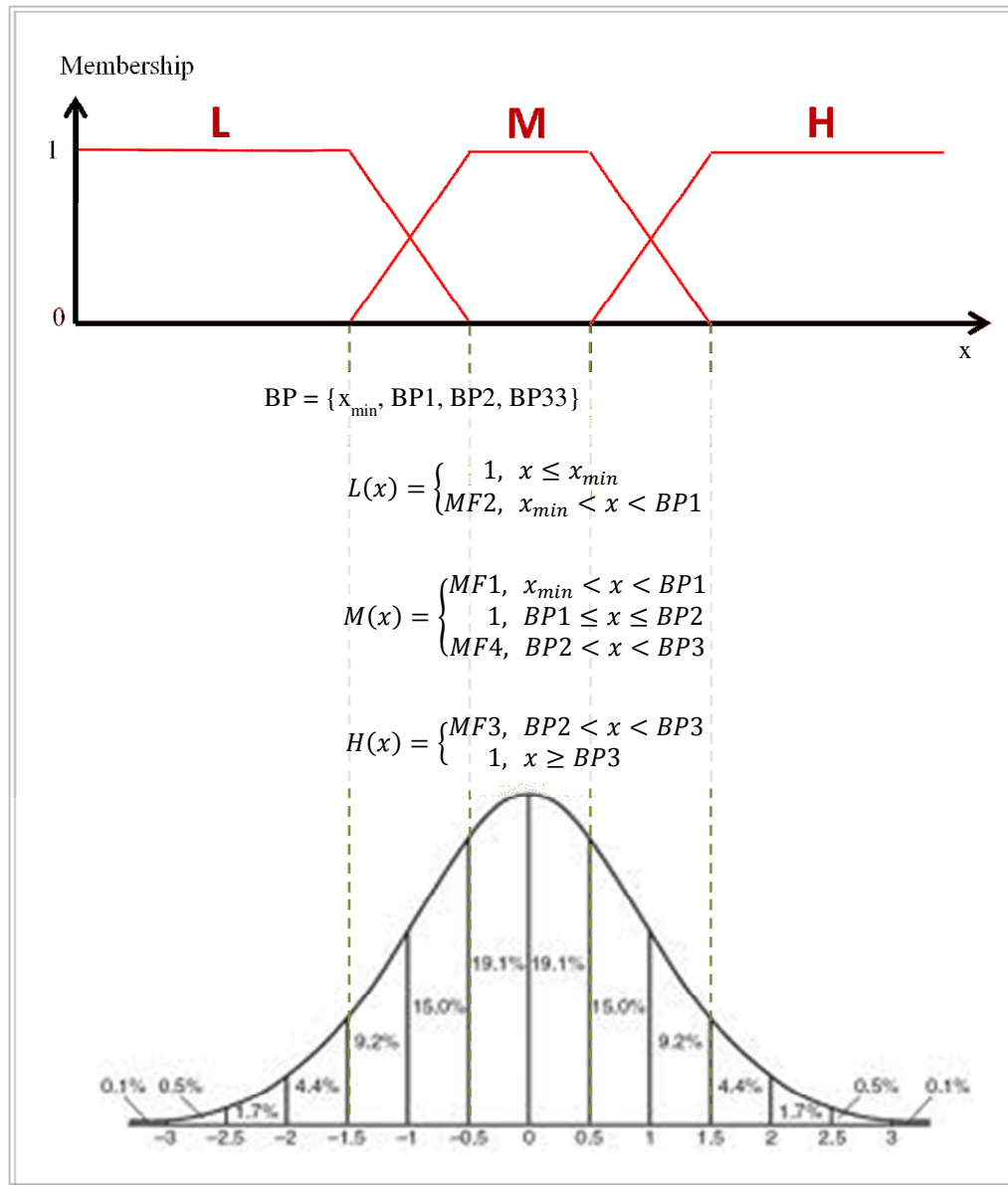


Figure 28: Normal distribution

5.4 Visual Analytics with Subjective Evaluation

SpatialARMED aims at proposing a visual analytic system to evaluate the interestingness of mined rules. The capability to support subjective evaluation and interactive visualization is an advantage of the framework. As proposed in Figure 29, there are two visual process pipelines for the SpatialARMED framework: one belongs to

the original spatial dataset and the other is for mined rules. During the first, geovisualization with mapping techniques assists the steps of spatial data exploratory analysis. This is particularly important in the processes of spatial autocorrelation analysis and spatial join. After spatial association rules are mined, the overwhelming quantity of them needs visual representation for assessment. The second visual pipeline thus is designed to handle this matter. Every step in these two visual processes involves interaction with the analyst for the purpose of selecting the most efficient visual representation, the most suitable evaluation criteria, and the most optimal subgroup selection, as well as for harvesting insights which could beneficially support the discovery of new and interesting rules. The system is developed as a stand-alone platform while providing rooms to either adopt or advance existing effective visualization techniques reviewed in Section 2.4.1. The 3D matrix AR visualization approach proposed in (Wong et al. 1999; Hofmann and Wilhelm 2001) is adopted here. This approach is not the most effective choice for SAR visualization in term of user interaction with dynamic rule selections, especially when dealing with a large number of rules. However, in this study, it works well under the SpatialARMED framework due to the proposed rule-subgroup evaluation scheme (which will be discussed in the following paragraphs). In addition, the visual attractiveness and simplicity of the 3D matrix visualization method create efficiency for the rule evaluation process. The availability of the open-source code also brings valuable advantages for inexperienced programmers who want the flexibility for potential customization.

While visual components and visual process pipelines shown in Figure 29 help create the machine-human interactive interface for analysis, the assessment mechanism of

SpatialARMED rule evaluation remains at the so-called *domain knowledge integrated rule evaluation process* proposed in Figure 30. Being designed to work with big data, the crucial components of this evaluation process are the library of known and unknown associations and interactive mechanism for subgrouping rules for visual analysis.

The library of known and unknown associations comprehensively represents the domain knowledge base which will be used later to subgroup and evaluate the mined rules. Known association are defined to be ones which are either well documented in the related literature or well acknowledged by a domain expert as having an association to the phenomenon under study. On the other hand, unknown association are the ones that are either not recognized or for which there exists a controversy as having an association to the phenomenon under study. Constructing the library of known and unknown associations is a challenging task that depends on the level of complexity in terms of integrated domain knowledge and the capability for expansion. For example, one could consider constructing a library of known and unknown associations which permits complex inter-association relationships (e.g. combination of different associations at different level of contributions) and/or allowing integration of knowledge from various sources or various domain experts. For this study, the simplest format is considered, which uses a relational table expressing knowledge of *known* (K) or *unknown* (U)

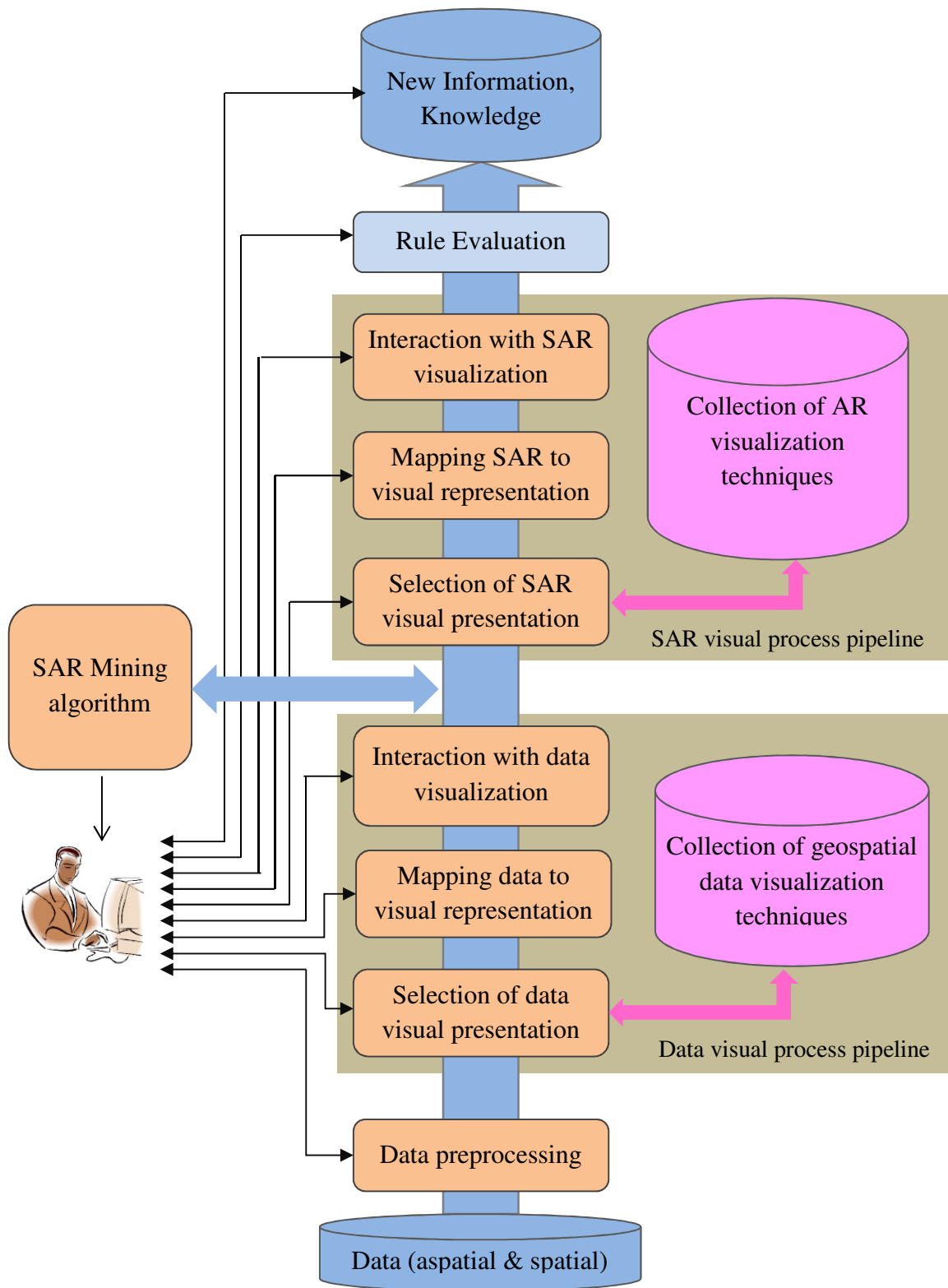


Figure 29: Pipeline and Components of Visual Analytic Process in SpatialARMED

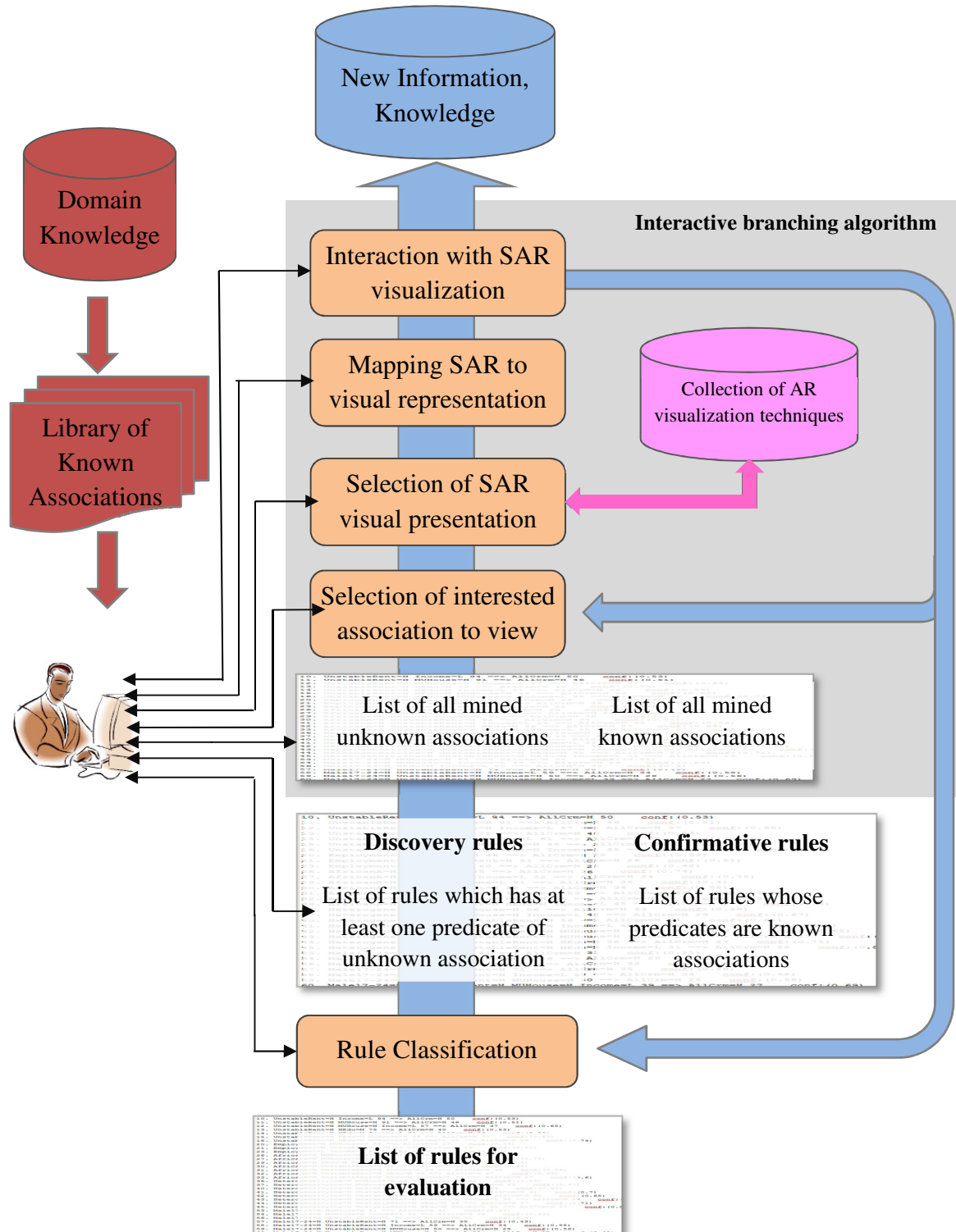


Figure 30: SpatialARMED domain knowledge integrated rule evaluation process using interactive branching approach

associations between phenomena to proposed associates. So the columns are for phenomena under study (e.g. crime) and the rows are for its associates.

Once the library of known and unknown associations has been established, it will be used in the rules evaluation process as the existing domain knowledge. Regarding a particular library, one could certainly question what is stated as known or unknown. However, with SpatialARMED, it is straightforward to modify or extend the library as needed and rerun the rules evaluation process for corresponding set of rules.

Using the established library of known and unknown associations, mined rules are classified into two categorized: *confirmative rules* and *discovery rules*, as shown in Figure 30. Confirmative rules are the rules whose all predicates are all known. The discovery rules are rules containing at least one unknown predicate. It is important to note that a discovering rule is not necessarily a new (i.e. interesting and useful) rule. Rather, it is a subject for further examination, i.e. still in its discovering process, because it contains an unknown predicate. Visual analytics on the set of confirmative rules mainly serve to evaluate the mining algorithm in confirming the known knowledge reported in the library of known and unknown associations, so-called *confirmative* in nature. Conversely, visual analytics on the set of discovery rules aim to discover interesting and potentially new rules, so-called *discovering* in nature.

The subsequent step in the mining process is to focus on the set of *discovery rules* for detecting potentially new and interesting ones. In typical cases, the number of discovery rules is substantial and, with big data, it becomes impossible to manually evaluate each and every discovering rule. In addition, most visualization methods face challenges to view all of these rules at once. The SpatialARMED framework proposes an approach,

called interactive branching evaluation approach, which subgroupes the discovering rules to support the discovery of new rules, as shown in Figure 30. One way to achieve this subgrouping procedure is: firstly to substract all unknown predicates involved in the set discovery rules; secondly, based on the domain knowledge embedded in the library of known and unknown associations, select promising unknown predicates in relation to the phenomenon under study which could potentially lead to the discovery of new interesting rules ; thirdly to select subset of discovering rules containing the promising predicate; and fourthly to evaluate these smaller subgroups of rules. This process is repeated until there is no further promising predicates identified. For example, assume one is mining SARs to *high crime*. *Low income* is a well-documented association to *high crime* while there exists some controversy over the association of *high income* to *high crime*. Thus, *low income* is defined as a *known* association and *high income* as an *unknown* (i.e. put into the discovery process) to *high crime*. Due to this defined library of known and unknown associations to high crime, all rules containing high income spatial predicates will be put into the pool of discovery rules. During the interactive branching evaluation process, the analyst can use high income as a particular promising predicate to select a subgroup of discovery rules for futher evaluation. The iterative branching evaluation process could even be repeated within subgroup of rules to assess groups of rules pertaining two or more promising predicates.

In comparison with traditional AR evaluation approaches which often apply top-down evaluation mechanisms and focus only on the few strongest rules, the interactive branching approach overcomes limitations due to the large number of rules while

allowing the integration of domain knowledge to increase the chance of discovering rare interesting rules.

CHAPTER 6: SPATIALARMED FOR CRIMINOLOGY

The performance of the SpatialARMED framework will be demonstrated for an application in criminology, aiming to mine spatial associations to crime in the City of Charlotte, NC. The mined results, i.e. rules, are compared to the current state of knowledge in spatial analysis of crime reviewed in Chapter 3 from both confirmative and novel perspectives.

6.1 Case Study – Dangerous Streets of High Criminal Activities

An experimental dataset of crime incidents obtained from the Charlotte-Mecklenburg Police Department (CMPD), including the UNC-Charlotte division, for 2010 is used for the mining task. The data set includes 67,595 point features representing the location of crime incidents of all types with XY coordinates and street addresses as shown in Figure 31. Non-criminal incidents reported to the police, such as missing person, suicide, overdose, sudden natural death, animal control issues, gas leak, vehicle recovery, fire and traffic events, are eliminated from the analysis. Details regarding criminal incident types and their shares within the dataset are shown in Table 3. To serve the purpose of the SpatialARMED framework demonstration and validation, this Chapter guides readers through the major steps to follow in order to mine spatial association rules with respect to crime, particularly crime of all types (CAT), motor vehicle thefts (MVT),

and thefts from motor vehicles (TFM). Motor vehicle related crime is chosen for mining in this particular case due to its high police reporting rate. In addition, visually distinguishable spatial patterns of these three different incident sets as shown in Figure 32 and Figure 33 provide a good context to examine the performance of SpatialARMED in term of discovering confirmative and new rules. Among 68,234 crime incidents of all types, 2,734 incidents are motor vehicle thefts and 8,152 are thefts from motor vehicles, under CMPD classification using National Incident Based Reporting System as reported in Table 4.

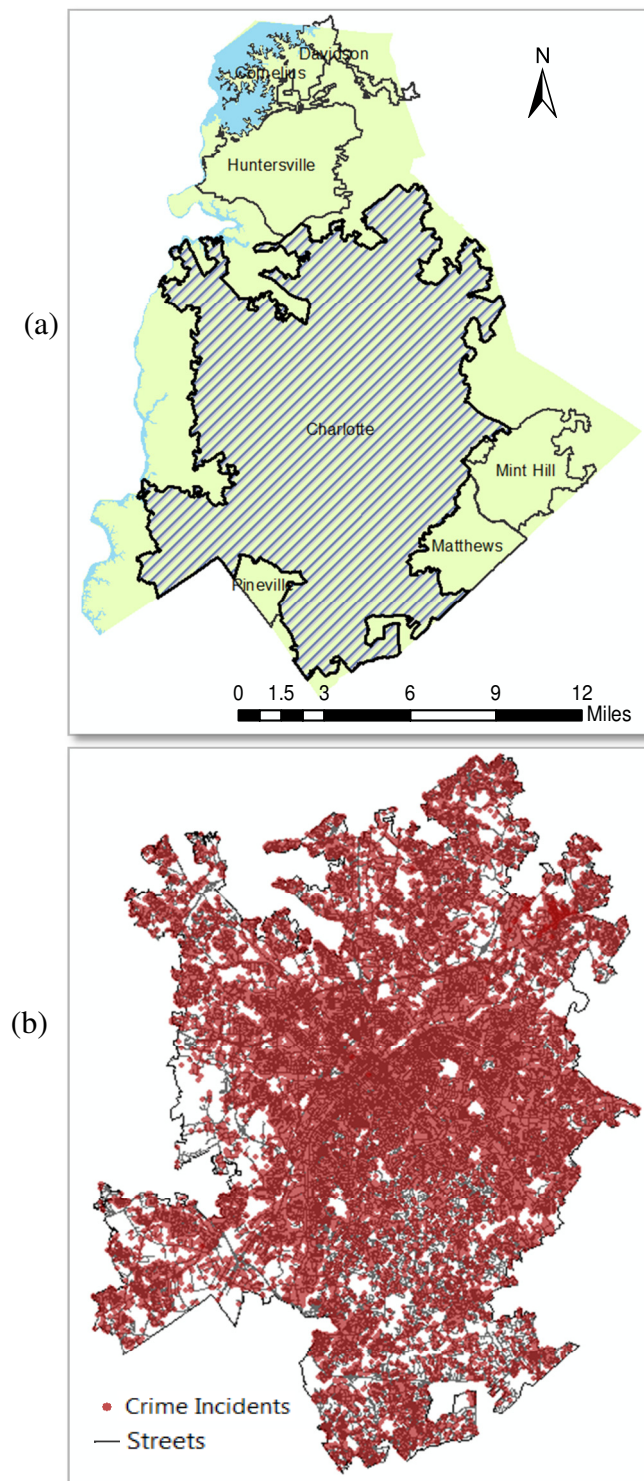


Figure 31: (a) Study area and (b) Charlotte street network and crime incidents

Table 3: Criminal Incident Count in Categories by CMPD National Incident Based Reporting System

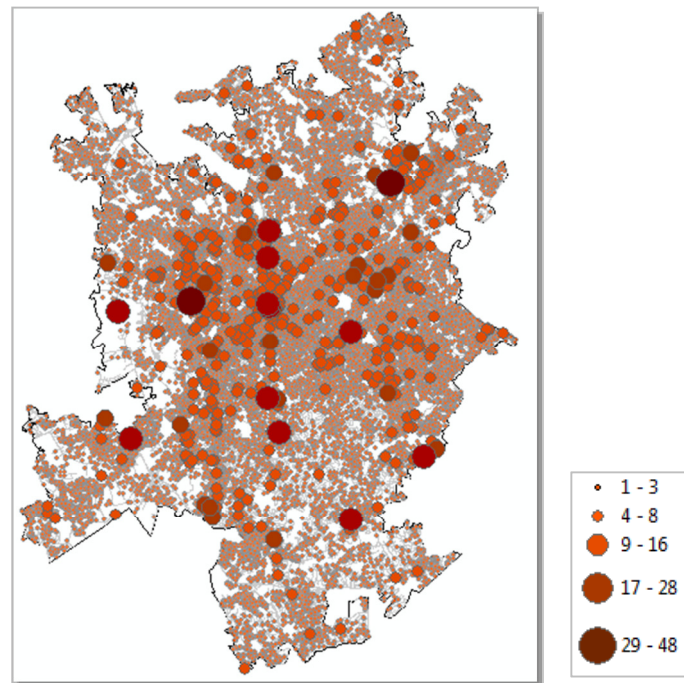
NIBRS categories	Count	%
ABC Violations / Liquor Law Violations	231	0.34
Animal control / Cruelty to Animals	10	0.01
Arson / Arson: Burning Church, Uninhabited House	25	0.04
Arson / Arson: Burning One's Own Dwelling	3	0.00
Arson / Arson: Burning Personal Property	91	0.13
Arson / Arson: Burning Public Building	4	0.01
Arson / Arson: Burning Schoolhouse	10	0.01
Arson / Arson: First & Second Degree	48	0.07
Arson / Arson: Mobile Home	1	0.00
Arson / Arson: Setting Fire to Woods, Grass, Field, Etc.	28	0.04
Assault / Assault: ADW Inflicting Serious Injury	207	0.30
Assault / Assault: ADW with Intent to Kill	45	0.07
Assault / Assault: ADW with Intent to Kill Inflicting Serious Injury	61	0.09
Assault / Assault: Aggravated Assault	318	0.47
Assault / Assault: by Pointing a Gun	306	0.45
Assault / Assault: Discharging Weapon into Occupied Property	105	0.15
Assault / Assault: Inflicting Serious Injury	48	0.07
Assault / Assault: on Child under 12	62	0.09
Assault / Assault: on Female	3050	4.47
Assault / Assault: on Government Officer or Employee	265	0.39
Assault / Assault: or ADW on Emergency Personnel	14	0.02
Assault / Assault: Simple	4119	6.04
Assault / Assault: Simple or Aggravated on Handicapped Person	27	0.04
Assault / Assault: with Deadly Weapon	704	1.03
Burglary / B/E: Felony Breaking or Entering	7751	11.36
Burglary / B/E: Misdemeanor Breaking or Entering	903	1.32
Burglary / Being Found Armed with Intent to Commit Burglary	1	0.00
Burglary / Burglary: First Degree	113	0.17
Burglary / Burglary: Second Degree	95	0.14
Burglary / Possession of Burglary Tools	9	0.01
Child Abuse / Child Abuse Inflicting Serious Injury	7	0.01
Child Abuse / Child Abuse: Misdemeanor	67	0.10
Child Sexual Assault / Child Abuse: Sexual Act	6	0.01
Child Sexual Assault / Rape: Forcible (against juvenile)	26	0.04
Child Sexual Assault / Sex Offense: First or Second Degree Forcible (child)	33	0.05
Child Sexual Assault / Sex Offense: Incest (involving child)	4	0.01
Child Sexual Assault / Sex Offense: Indecent Liberties with Child	93	0.14
Child Sexual Assault / Sex Offense: Sexual Activity by Custodian	1	0.00

Child Sexual Assault / Sex Offense: Sexual Activity by Substitute Parent	1	0.00
Child Sexual Assault / Sex Offense: Statutory Rape	35	0.05
Child Sexual Assault / Sex Offense: Statutory Sex Offense	9	0.01
Child Sexual Assault / Sexual Battery (Child)	47	0.07
Communicating Threats / Bomb Threat	22	0.03
Communicating Threats / Communicating Threats (against person)	2796	4.10
Communicating Threats / Communicating Threats (against property)	168	0.25
Damage to Property / Damage or B/E Coin or Currency-Operated Machine	26	0.04
Damage to Property / Injury to Real or Personal Property (including Graffiti)	6463	9.47
Embezzlement / Embezzlement/Larceny by Employee	281	0.41
Forgery / Forgery: Counterfeiting	483	0.71
Forgery / Forgery: Uttering Forged Checks/Documents/Notes	381	0.56
Fraud / Fraud: Financial Identity Fraud	828	1.21
Fraud / Fraud: Defrauding Innkeeper/Defrauding Taxi Driver	83	0.12
Fraud / Fraud: False Pretenses	892	1.31
Fraud / Fraud: Financial Transaction Card Withholding, Forgery or Fraud	1539	2.26
Gambling / Gambling	1	0.00
Homicide / Manslaughter: Voluntary or Involuntary	1	0.00
Homicide / Murder: First or Second Degree	32	0.05
Kidnapping / Kidnapping: False Imprisonment	22	0.03
Kidnapping / Kidnapping: Felonious Restraint	22	0.03
Kidnapping / Kidnapping: First or Second Degree	86	0.13
Larceny / B/E Vehicle (larceny from)	7743	11.35
Larceny / Damage or B/E Coin or Currency-Operated Machine	44	0.06
Larceny / Financial Transaction Card Theft	285	0.42
Larceny / Larceny: Felonious	2484	3.64
Larceny / Larceny: Misdemeanor	9340	13.69
Larceny / Shoplifting: by Concealment	696	1.02
Larceny / Shoplifting: by Price Switching	6	0.01
Miscellaneous criminal / Bribery	2	0.00
Miscellaneous criminal / Crime against Nature (Consensual)	31	0.05
Miscellaneous criminal / Extortion/Blackmail	10	0.01
Miscellaneous Criminal / Failure to Return Rented Property	12	0.02
Miscellaneous criminal / Family Offense: Non-violent (Abandon/Negligence/Etc)	102	0.15
Miscellaneous criminal / Harassing Phone Calls	1331	1.95
Miscellaneous criminal / Other Unlisted Criminal Offense	1743	2.55
Miscellaneous criminal / Possession of Stolen Goods: Misdemeanor or Felony	263	0.39

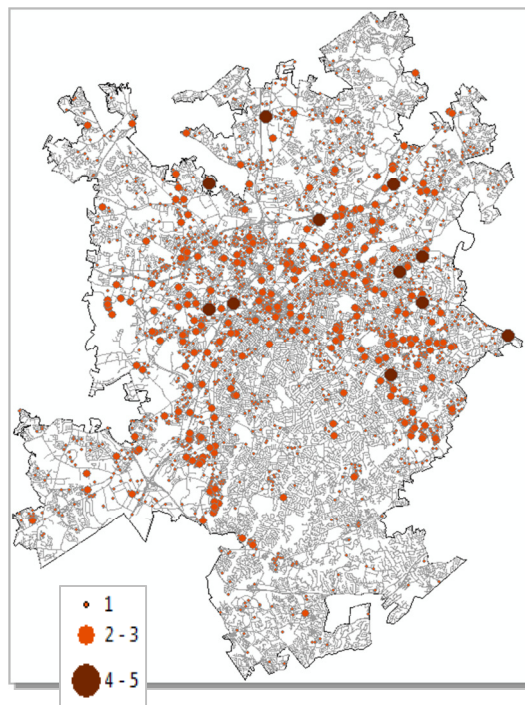
Miscellaneous criminal / Stalking	64	0.09
Miscellaneous criminal / Violation of Restraining Order	180	0.26
Miscellaneous criminal / Weapon Law Violation	796	1.17
Miscellaneous criminal / Worthless Check: Felony	39	0.06
Miscellaneous non-criminal / Obscene Phone Call	2	0.00
Narcotics / Drug Equipment Violation	1861	2.73
Narcotics / Drug/Narcotic Violation	2362	3.46
Narcotics / Drug/Prescription Fraud	53	0.08
Public disorder / Affray (fighting)	237	0.35
Public disorder / Curfew	49	0.07
Public disorder / Disorderly Conduct	134	0.20
Rape/Sex Offense / Rape: Forcible (against adult)	73	0.11
Robbery / Robbery: Armed	688	1.01
Robbery / Robbery: Common Law	175	0.26
Sex Offense/Assault / Indecent Exposure	48	0.07
Sex Offense/Assault / Peeping Tom	20	0.03
Sex Offense/Assault / Sex Offense: First or Second Degree Forcible (Adult)	27	0.04
Sex Offense/Assault / Sexual Battery (Adult)	43	0.06
Traffic / Death by Vehicle: Felony/Misdemeanor (Traffic Fatality)	4	0.01
Traffic / Driving Under the Influence	103	0.15
Traffic / Hit and Run (Personal Injury)	228	0.33
Trespass / Trespass: Domestic Criminal	13	0.02
Trespass / Trespass: First or Second Degree	540	0.79
Trespass / Trespass: Forcible	3	0.00
Vehicle Theft / B/E Vehicle (vehicle theft)	231	0.34
Vehicle Theft / Failure to Return Rented Vehicle	129	0.19
Vehicle Theft / Larceny of Vehicle	2302	3.39
Vehicle Theft / Unauthorized Use of Motor Vehicle	437	0.64
Vice / Pornography/Obscene Materials	4	0.01
Vice / Prostitution	169	0.25
Vice / Prostitution: Assisting, Promoting, Etc.	9	0.01

Table 4: Total incident count by crime of all types and motor vehicle related types by National Incident Based Reporting System (using the Highest Class)

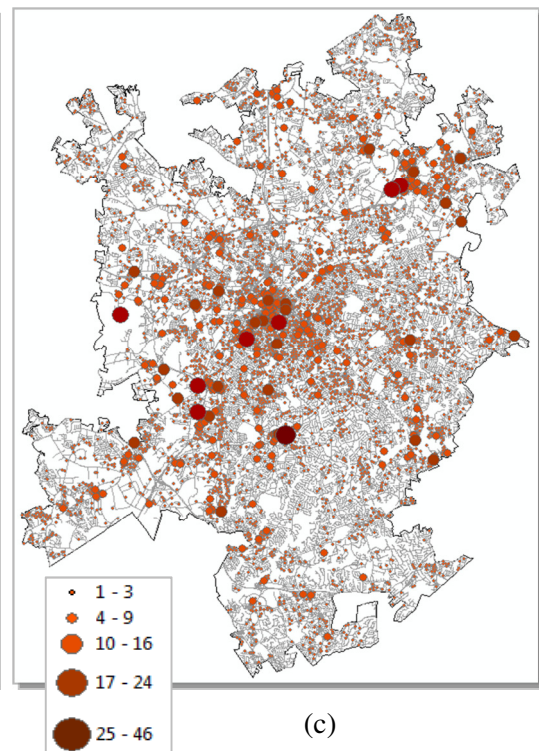
Criminal incident type	Count	%
All types	68,234	100
Motor vehicle theft	2,734	4
Theft from motor vehicle	8,152	12



(a)

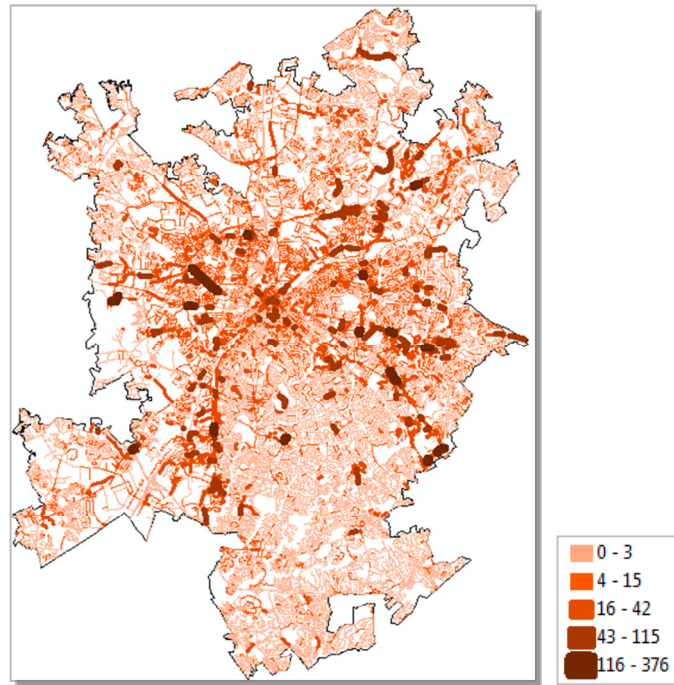


(b)

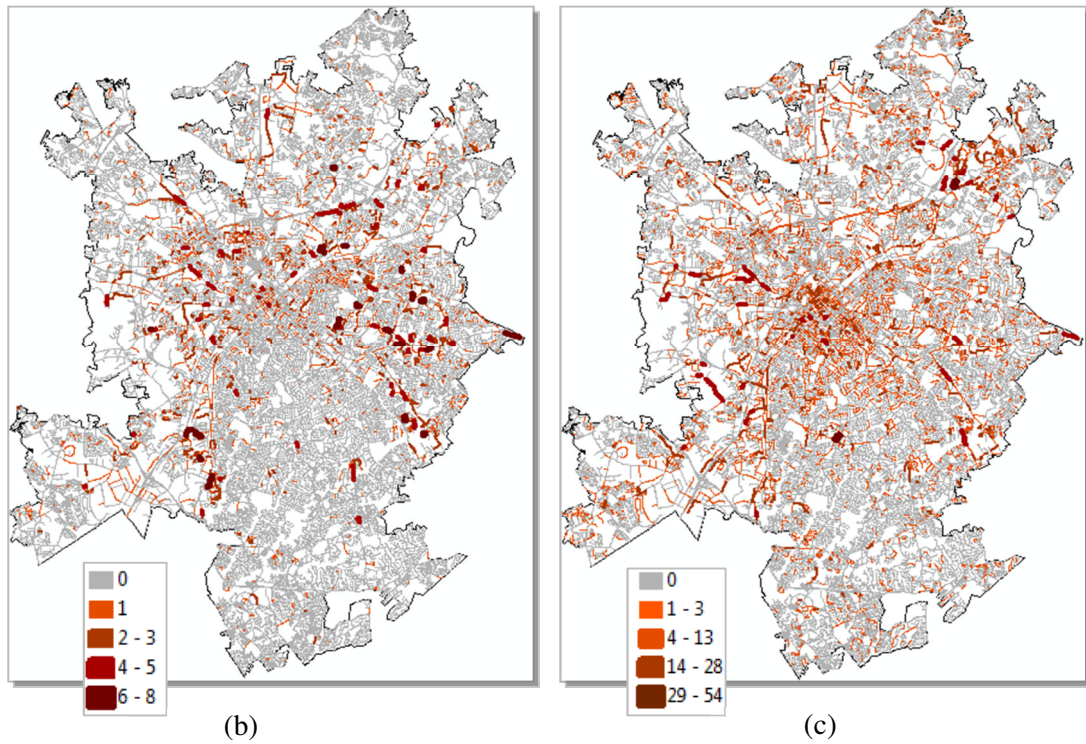


(c)

Figure 32: Incident count on occurrence locations of (a) crime of all types (CAT), (b) vehicle theft (MVT), and (c) theft from vehicle (TFM)



(a)



(b)

(c)

Figure 33: Incident count on associated street blocks of (a) crime of all types (CAT), (b) motor vehicle theft (MVT), and (c) theft from motor vehicle (TFM)

Regarding datasets for task-relevant variables, i.e. potential associations to crime of all types or motor vehicle related types, the following sources of data generate the associates (more detail is shown in Table 5 and Figure 34): American Community Survey (ACS) five-year estimates for 2005-2010 at the block group level, census 2010 data at the census block level, and ancillary data such as business data (i.e. locations of retail stores, restaurants, service offices, etc.), Wal-mart super center locations, alcohol drinking places, hotels and motels, shopping malls, and park-n-ride facilities. A list of reference variables and task related variables (i.e. potential associates) are identified. These are reported in Table 6. Among these variables, the heterogeneity index is estimated as $(1 - \sum P_i^2)$, where P_i is the percentage of population in decimal value for a particular racial group.

As crime incidents are recorded to street addresses, it is reasonable to consider the distribution of crime being constrained to the street network. Street blocks, i.e. the street segments on which one could travel without crossing any other streets, are thus used as spatial dependence structure analysis units for crime variables.

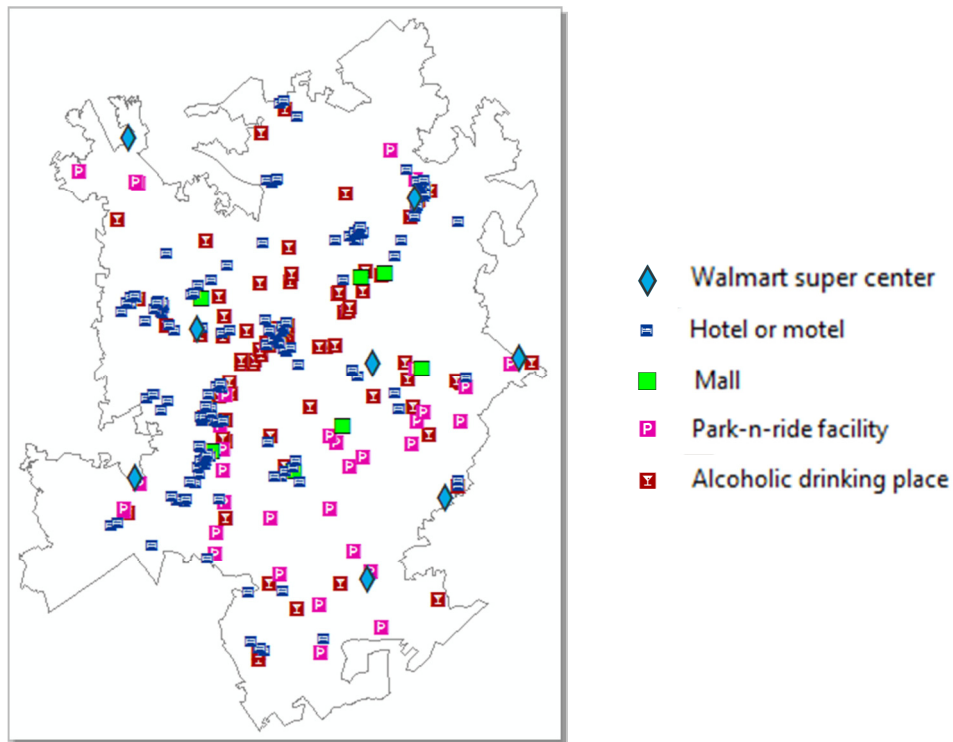


Figure 34: Representative facilities as potential crime generators and attractors in the City of Charlotte

Table 5: Data sets

	Attributes	Source	Year	Type
1	Crime incident	CMPD	2010	Point
2	Population	Census	2010	Areal BLK
3	Race	Census	2010	Areal BLK
4	Sex and age	Census	2010	Areal BLK
5	House tenure	Census	2010	Areal BLK
6	Family types	Census	2010	Areal BLK
7	Income	ACS	2005-2010	Areal BLG
8	Education	ACS	2005-2010	Areal BLG
9	Employment	ACS	2005-2010	Areal BLG
10	Multiple unit housing	ACS	2005-2010	Areal BLG
11	Population by move in year, by home tenure	ACS	2005-2010	Areal BLG
12	Retail business locations	MECK	2008	Point
13	Park-and-ride facilities	CATS	2009	Point
14	Shopping malls	MECK	2010	Point
15	WAL-MART super centers	MECK	2010	Point
16	Hotel and motels	MECK	2010	Point
17	Alcoholic drinking places	MECK	2010	Point

Table 6: Identified variables for spatial dependence structure quantification

Attributes	Resolution	Dataset
Incident count: Crime of all types	Street block	Crime
Incident count: Motor vehicle thefts	Street block	Crime
Incident count: Thefts from motor vehicle	Street block	Crime
Business count	Street block	Business
Per capital income	Block group	Income
% of population with high school degree or higher	Block group	Education
% of population employed	Block group	Employment
% of housing units in multiple (>3) unit structures	Block group	MU housing
% of population who rented and moved in <5 years (also referred to as “unstable rent” in this study)	Block group	Rent pop. by move-in year
Heterogeneity index	Census block	Race
% of African American population	Census block	Race
% of males 17-28 population	Census block	Sex and age
% of owner-occupied homes	Census block	House tenure
% of single-parent families	Census block	Family type

6.2 SpatialARMED Implementation Aspects

The SpatialARMED framework encompasses several loose-coupled algorithms and procedures which must be operated in a chronological sequence. Although this configuration could potentially encounter some limitations for applications that prioritize automation and timely responses, it offers implementation options to handle complexity and allows human interactions during the process of mining. Figure 35 summarizes and illustrates the implementation workflow of the SpatialARMED framework for the case study analysis. The associated hardware, software, and procedures for each analysis level

are outlined. Generally, ESRI ArcGIS 10.0 was used for data preparation, data manipulation, mapping, and early exploratory data analysis.

In the second level of SpatialARMED methodology, which deals with spatial analysis, a significant portion of tasks consists in cluster detection for crime and its associates. The parallel AMOEBA algorithm presented in Section 5.2.1 was used due to the complexity of the spatial dependency structures embedded in the data. The UNC-Charlotte parallel high performance computing (HPC) clusters were used to facilitate this process. Generally, the UNC-Charlotte HPC center operates in a Redhat Linux based high performance computing environment that includes HPC clusters and systems of various capabilities serving a variety of campus research communities. This research utilized the VIPER and GEM clusters in particular due to their availability for access. The VIPER cluster encompasses 88 computing nodes with 840 computing cores and 39 TBs usable RAID storage (96 TBs raw), serving as a general use cluster in any faculty sponsored research project. The GEM cluster, on the other hand, is a small cluster dedicated to Geospatial Modelling in the UNC-Charlotte Center for Applied Geographic Information Science (CAGIS) which involves only 2 computing nodes with 64 cores and 64 GBs RAM. Using the P-AMOEBA approach for this study, the AMOEBA algorithm implemented in the open source Python-based software package called clusterPy by RiSE (Research in Spatial Economic) group (Duque et al. 2011) was customized for use with parallel processing. Data was decomposed into 100 portions following the regional decomposition scheme presented in Figure 23 and distributed to 100 computing cores of VIPER and GEM clusters for ecotope identification for each cell. The recompilation of parallel output and detection of clusters are then done as a separated second phase,

following the workflow presented in Figure 22. This process is repeated for every variables involved in the study. While units of analysis for the data of census or ACS sources are areal unit equivalent to census blocks or block groups, those for crime incidents and retail stores are the corresponding street blocks. As the final result, spatial dependence structures represented by clusters of high or low values for each and every variable are identified.

The next step in the SpatialARMED spatial analysis component requires modeling the spatial spillover effect of the hot and cold clusters as elaborated in Section 5.2.2 (Equations 3, 4, and 5). Python-based and VBA-based programming modules in ESRI ArcGIS 10.0 were developed by the author to implement this task. The most expense in term of computing resource during this task remains the estimation of the from-to cost matrix in term of distance for every analysis units of BLK, BLG, or street blocks to hot/cold cluster elements in the study area. Given the big data scheme, this requirement could imply a challenge for run time improvement. For this particular case study, the distances are preestimated, saved into text files and recalled as necessary when running the spatial spillover modeling algorithm.

In the next levels of SpatialARMED, including predication, rule mining, rule evaluation and visualization, implementation are based on Python-based stand-alone programming developed by the author (e.g. to serve crisp and fuzzy predicate mapping or rule classification) or on Java-based customization of existing open source software packages including ARM mining package from the Liverpool University – Computer Science Group (LUCS) (Brin et al. 1997b; Coenen 2004) and ARV (Association Rule Viewer) software by the Computer Science Laboratory of Lille (d'Informatique

Fondamentale de Lille) in France (Open source downloadable at <http://www.lifl.fr/~jourdan/download/arv.html>). For the case of LUCS ARM software, customization was necessary to deal with modifications in input/output format and rare rule issues. For ARViewer, customization of the software served the purpose of accommodating extra procedures to convert the rule input from text format to XML format.

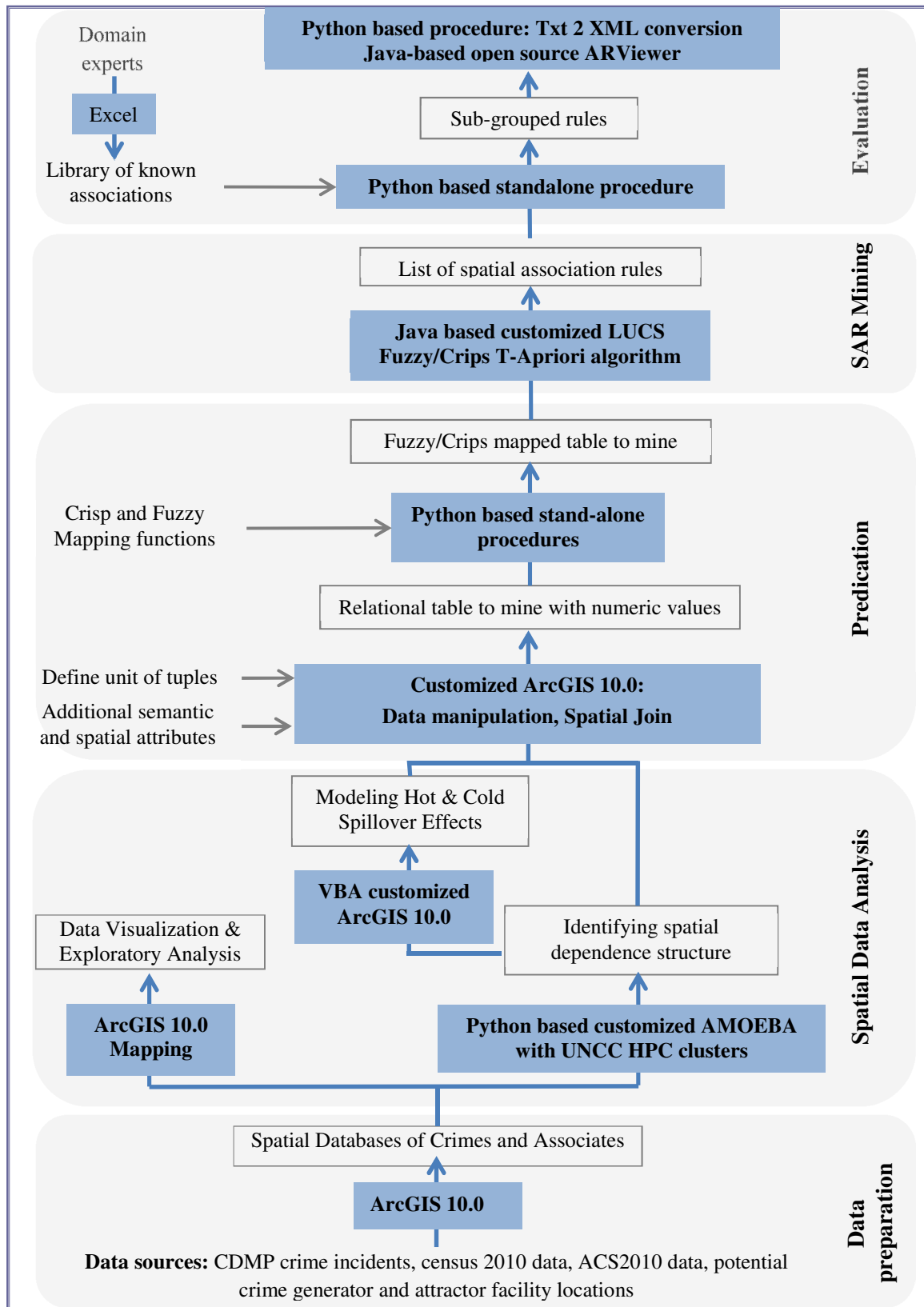


Figure 35: Spatial implementation work flow for case study

The following subsections illustrate and discuss the application of the SpatialARMED framework with results at each of its levels while mining spatial associations to crime of all types combined and also motor vehicle related crimes, separately.

6.3 SpatialARMED Level 2: Spatial Data Analysis

6.3.1 Identifying Spatial Dependence Structures

As discussed in Chapter 5, the first important task within the SpatialARMED framework is to identify the spatial dependence structure of crime and its associated variables using the AMOEBA algorithm. AMOEBA spatial clustering is performed to identify the so-called dangerous streets due to CAT, MVT, and TFM, which are highlighted in red and shown in Figure 36, Figure 37, and Figure 38, respectively. A very distinguishable spatial pattern of CAT concentration is observed in central Charlotte, along the major road corridors, and around the University of North Carolina at Charlotte (UNC-Charlotte) campus area. South Charlotte, on the other hand, is shown to have low CAT. Motor vehicle thefts possess a spatial pattern similar to CAT. It is, however, interesting to see the spatial pattern of TFM, which appear to be spread through the entire study area, with no zone of high concentration.

AMOEBA spatial clustering is also performed for every associate variable listed in Table 6 with its highest possible spatial resolution; results are depicted in Figure 39 to Figure 49. This process aims to identify clusters, i.e. hot and cold spots, for each variable. The spatial clustering information generated by the AMOEBA algorithm for each variable includes the number of clusters and their related details such as G^* values and cell members. Clusters of positive G^* values represent hot spots and those of negative values represent cold spots. A common color scheme is used across these figures: red

color is used to represent clusters of G high value (i.e. hot spots) while a blue is for clusters of low G value (i.e. cold spots). The AMOEBA algorithm is proven to be robust in terms of detecting spatial clusters due to its capability to conduct multi-directional searches for irregular clusters at the highest possible resolution. One of the major advantages of using AMOEBA is that the identified clusters are detected from the data as the result of the search for ecopoles and of the Monte Carlo based significance test for the final non-overlapping clusters of highest G values. With this algorithm, the definition of “high” or “low” clusters are mined and quantitatively defined. During the process of generating spatial predicates for SAR mining, these clusters will be used to establish meaningful spatial relations such as *within high/low concentration*, *near-by high/low concentration*, or *under the influence of high/low concentration*, etc. From mining perspectives, this spatial cluster mining process helps to eradicate limitations due to predetermined concepts of spatial patterns and spatial processes.

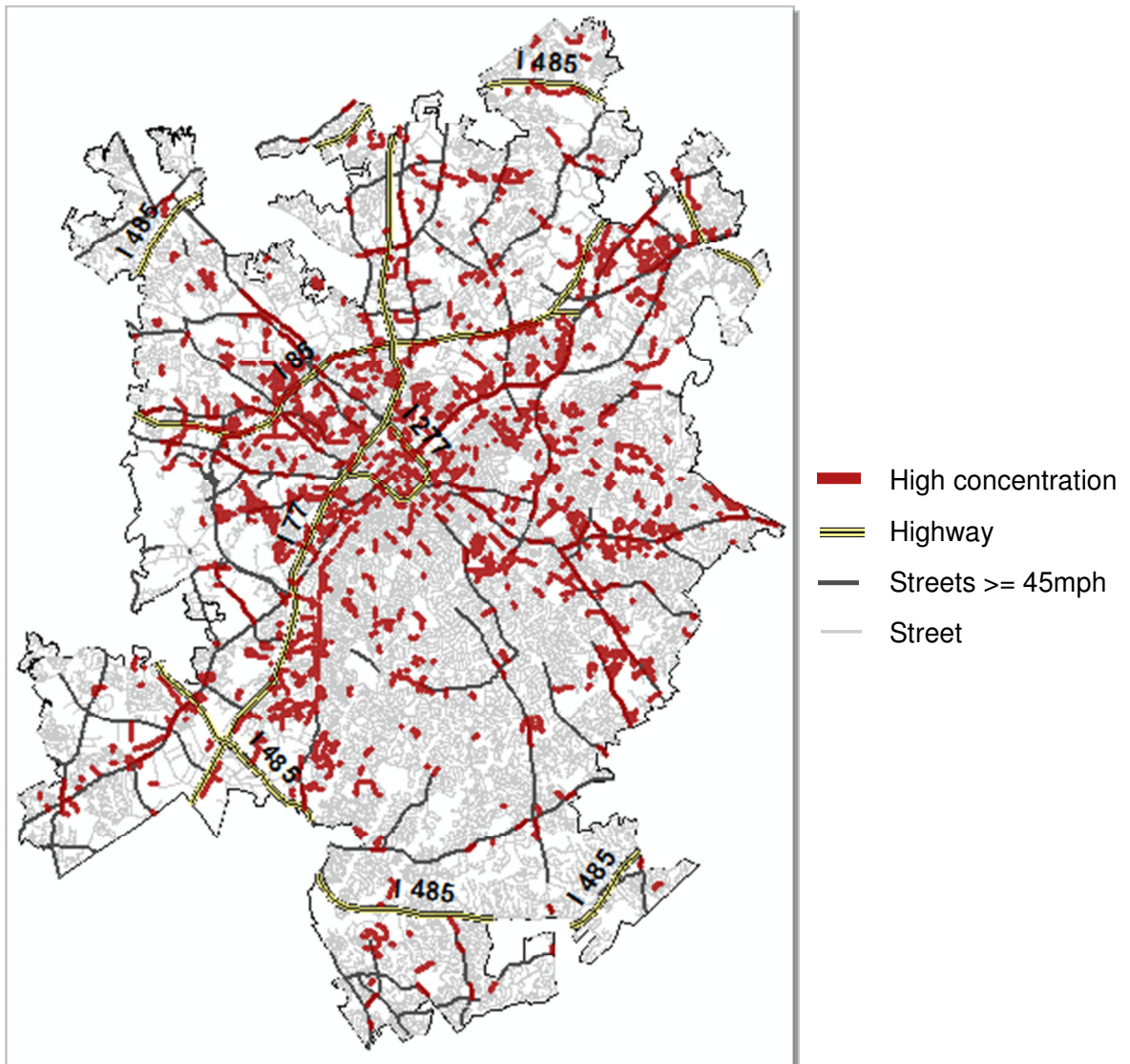


Figure 36: AMOEBA-based hot spots for crime of all types (CAT)

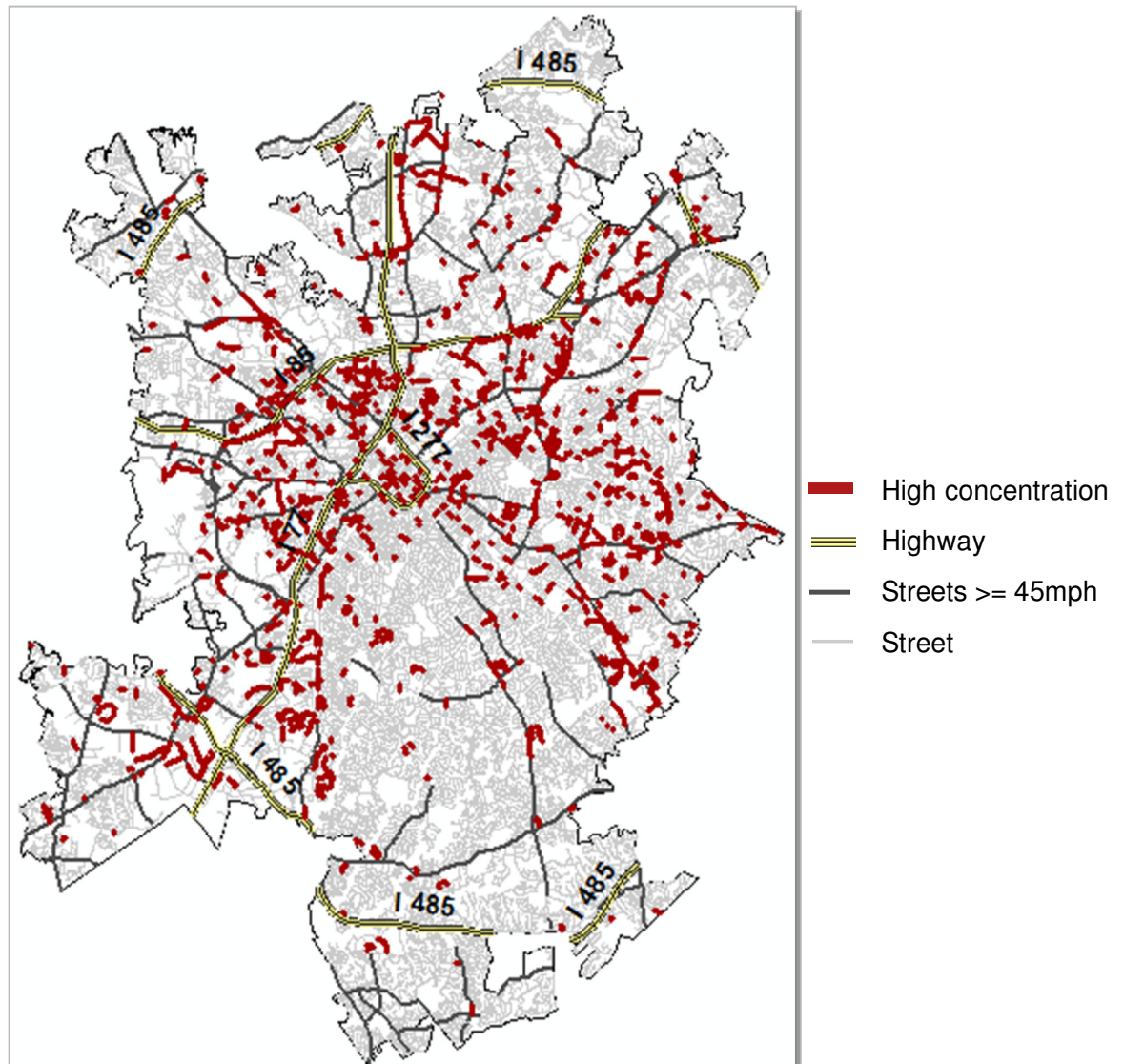


Figure 37: AMOEBA-based hot spots for motor vehicle thefts (MVT)

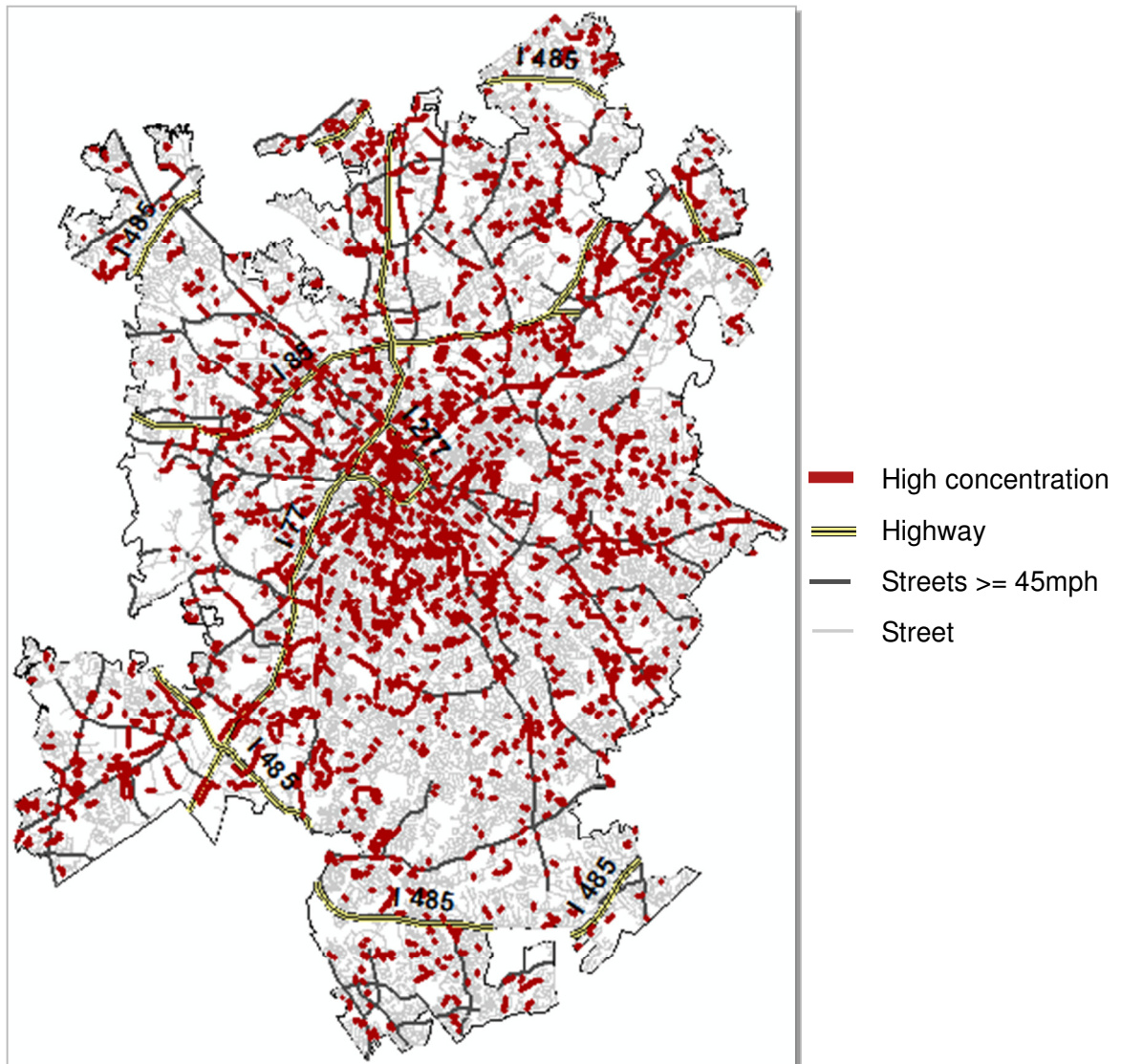


Figure 38: AMOEBA-based hot spots for thefts from motor vehicle (TFM)

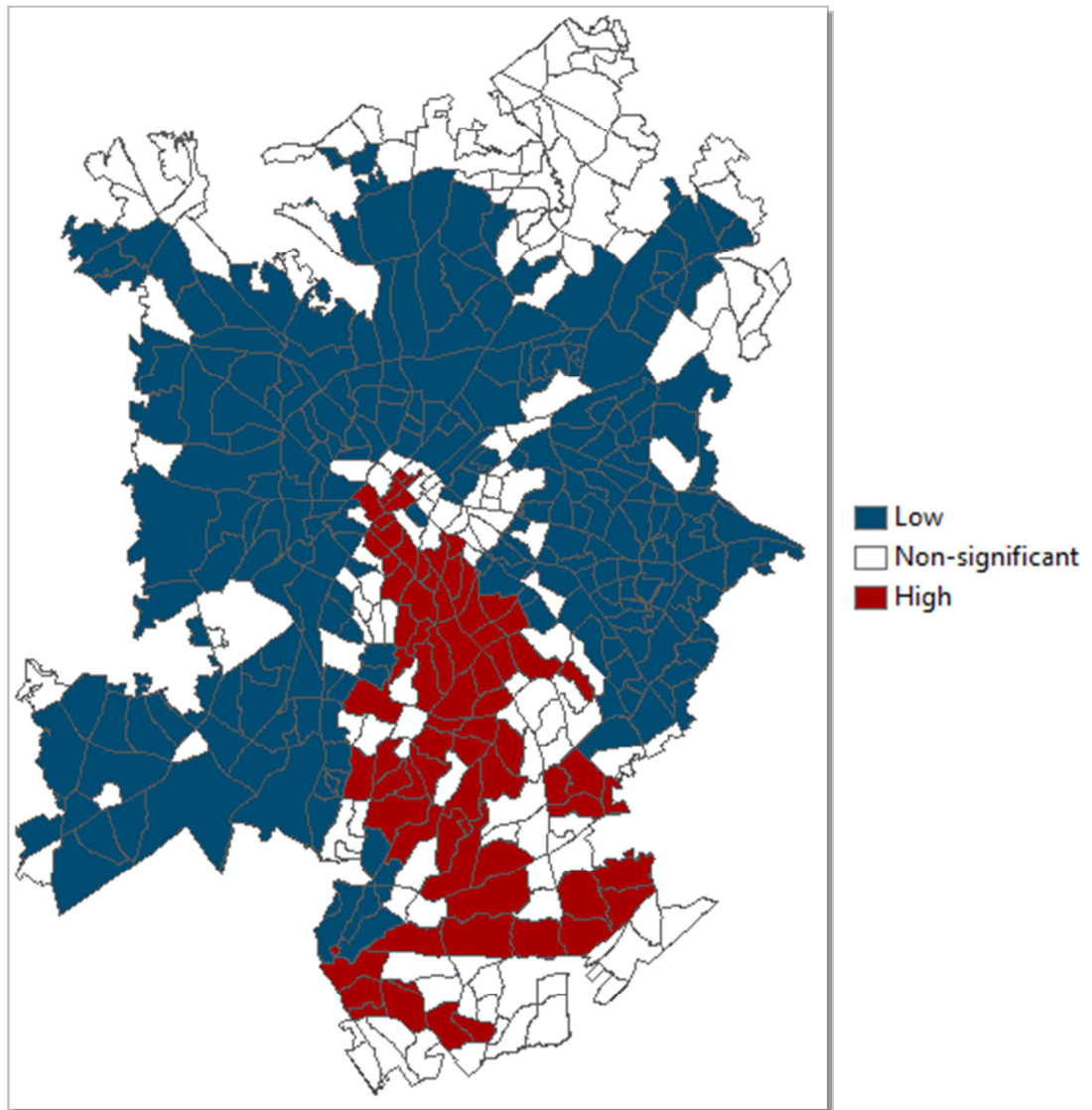


Figure 39: AMOEBA-based clusters for per capital income using ACS2010 BLG data with Rook neighborhood definition

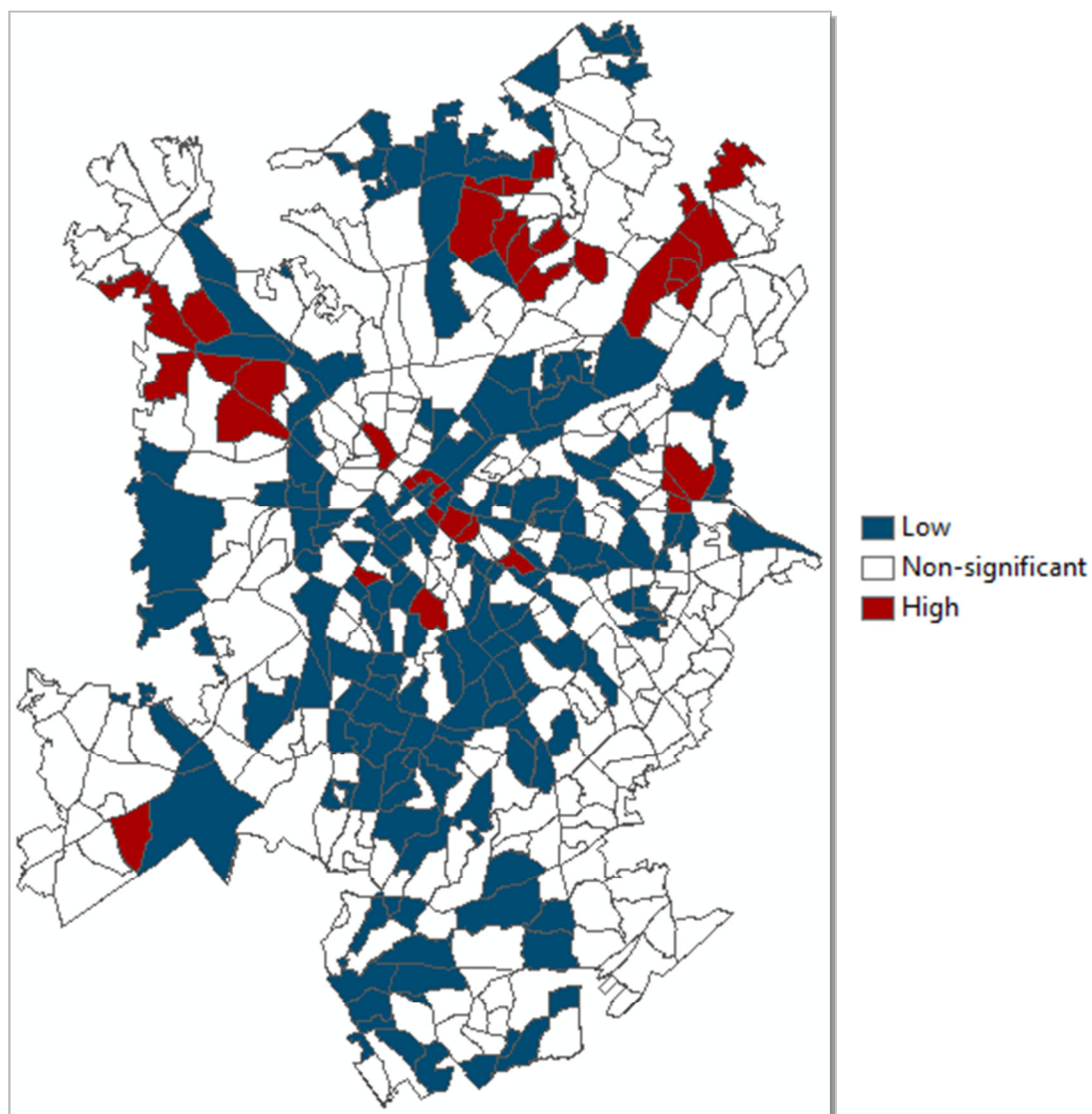


Figure 40: AMOEBA-based clusters for percentage of population with high school degree or above using ACS2010 BLG data with Rook neighborhood definition

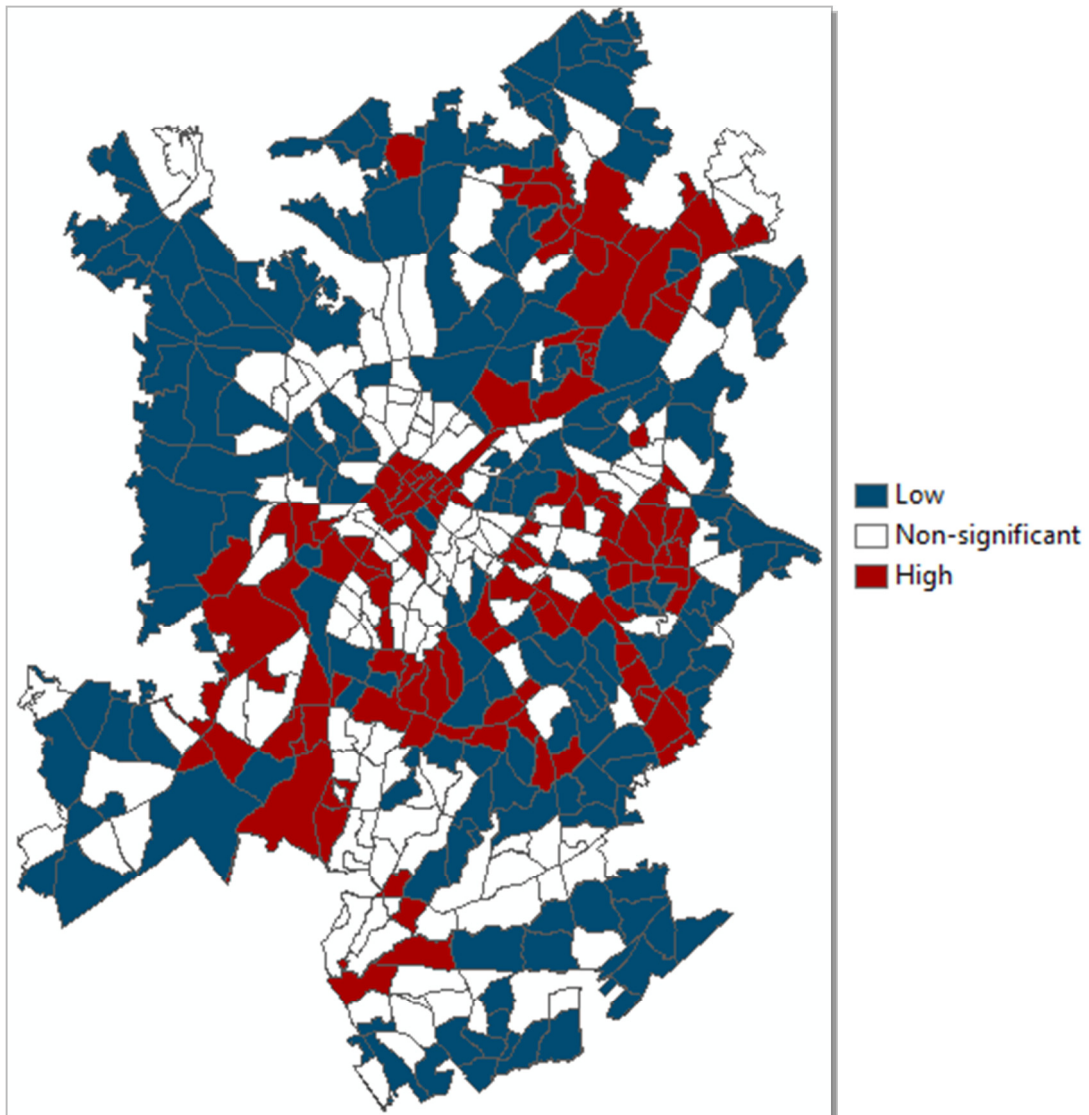


Figure 41: AMOEBA-based clusters for percentage of multiple (> 2) unit homes using ACS2010 BLG data with Rook neighborhood definition

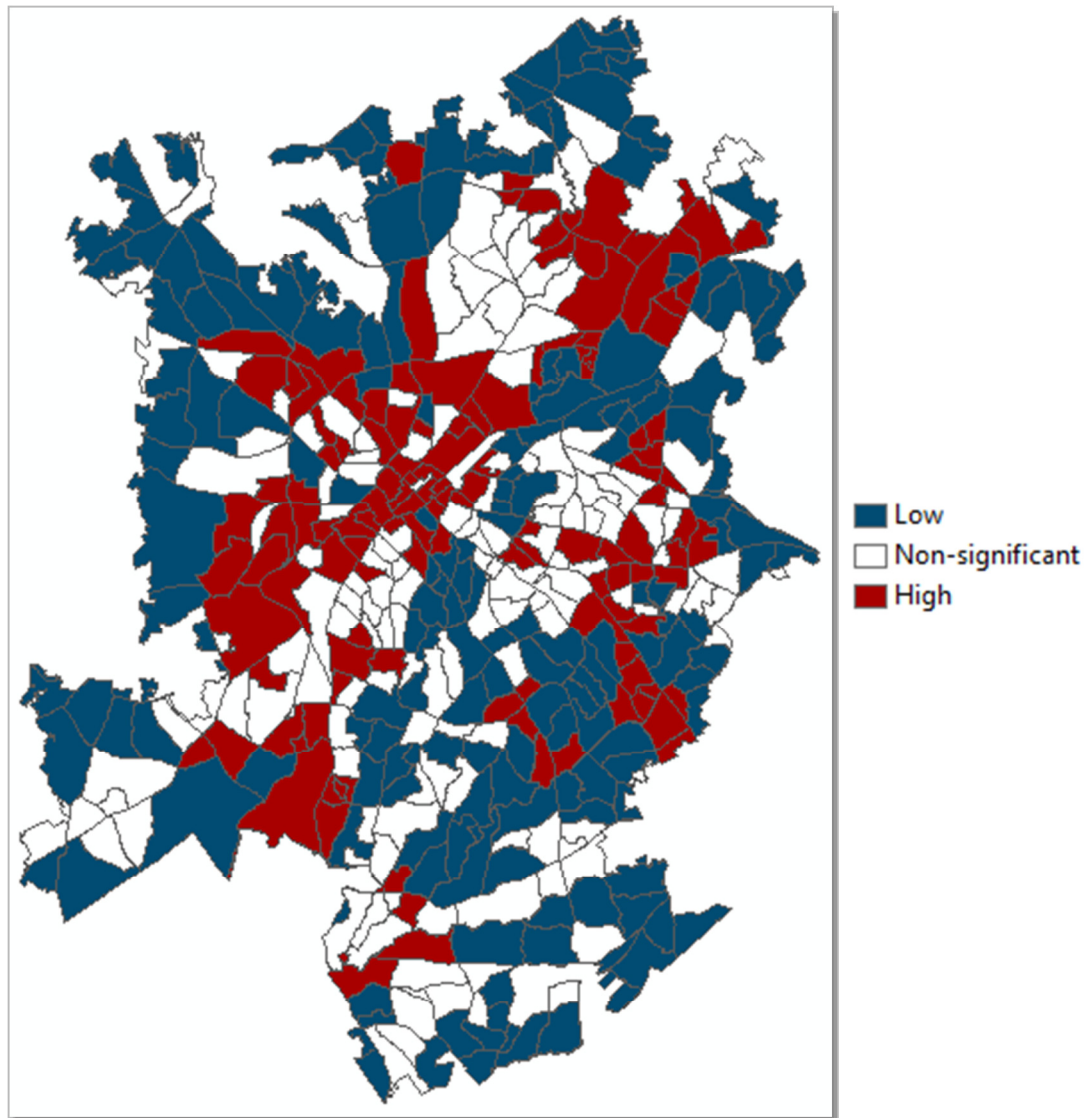


Figure 42: AMOEBA-based cluster for percent of population who rent and move in less than 5 years using ACS2010 BLG data with Rook neighborhood definition

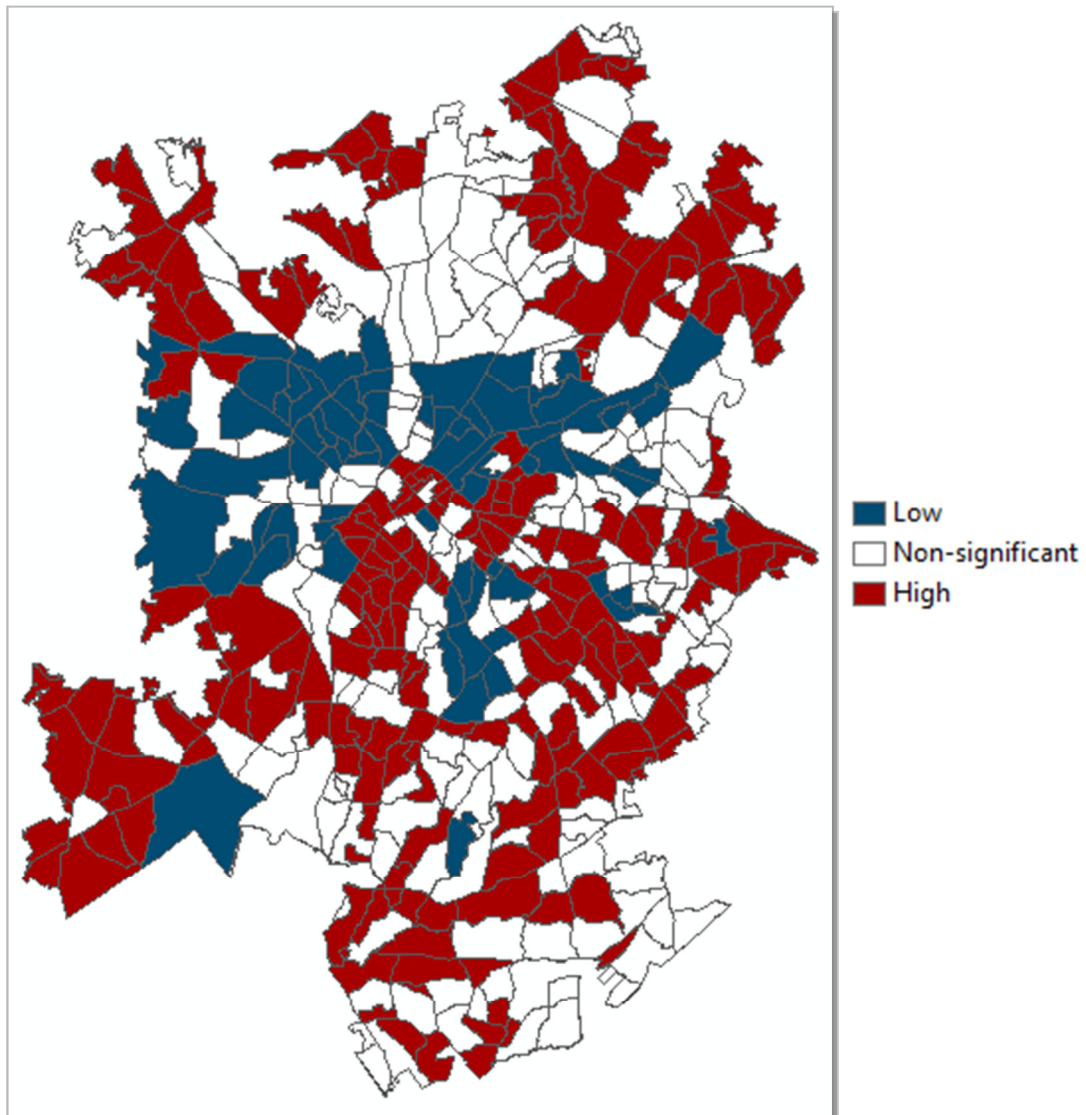


Figure 43: AMOEBA-based clusters for percentage of employed population using ACS2010 BLG data with Rook neighborhood definition

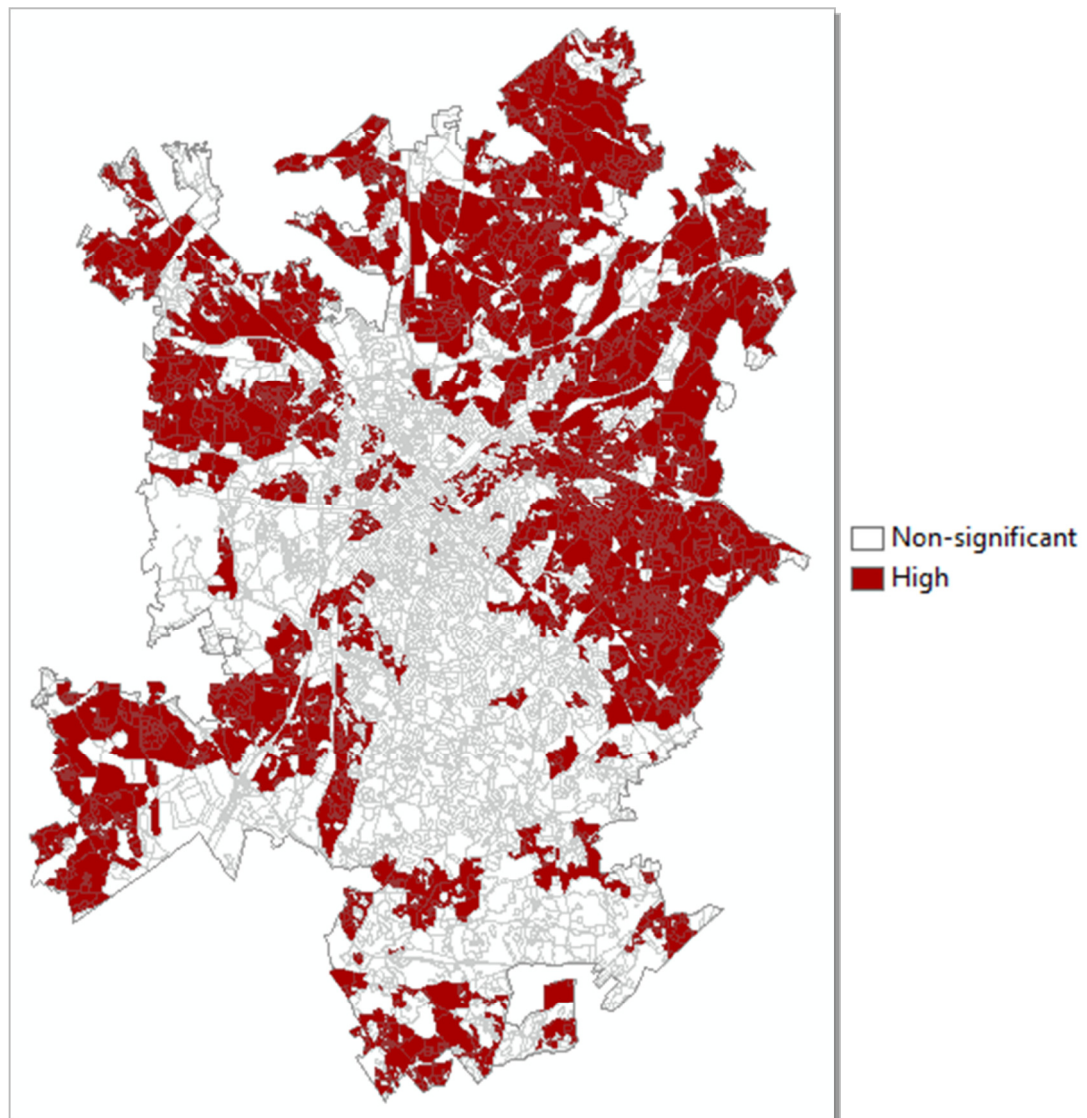


Figure 44: AMOEBA-based clusters for ethnic heterogeneity index using Census 2010 BLK data with Rook neighborhood definition

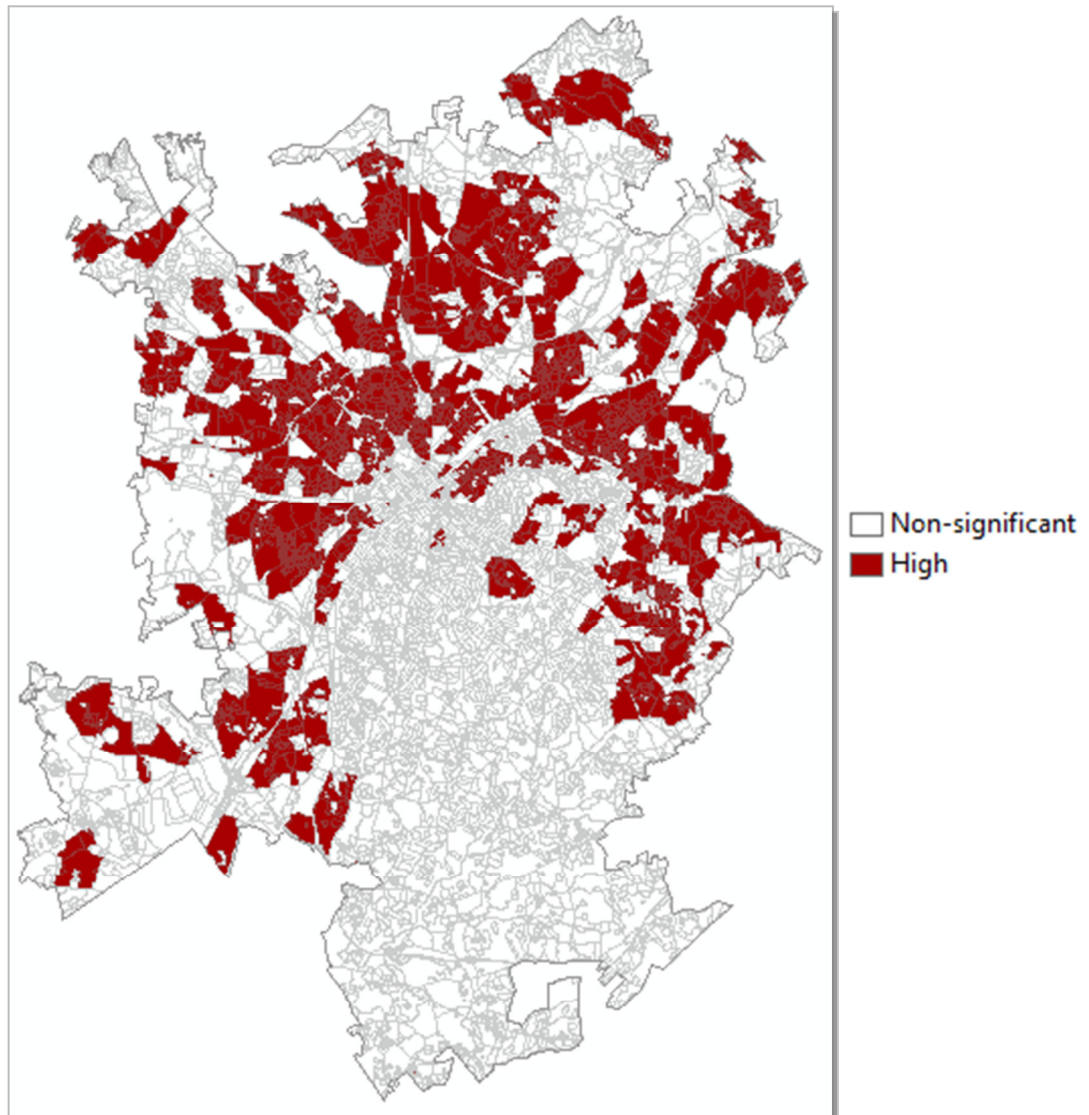


Figure 45: AMOEBA-based clusters for percentage of African-American population using Census 2010 BLK data with Rook neighborhood definition

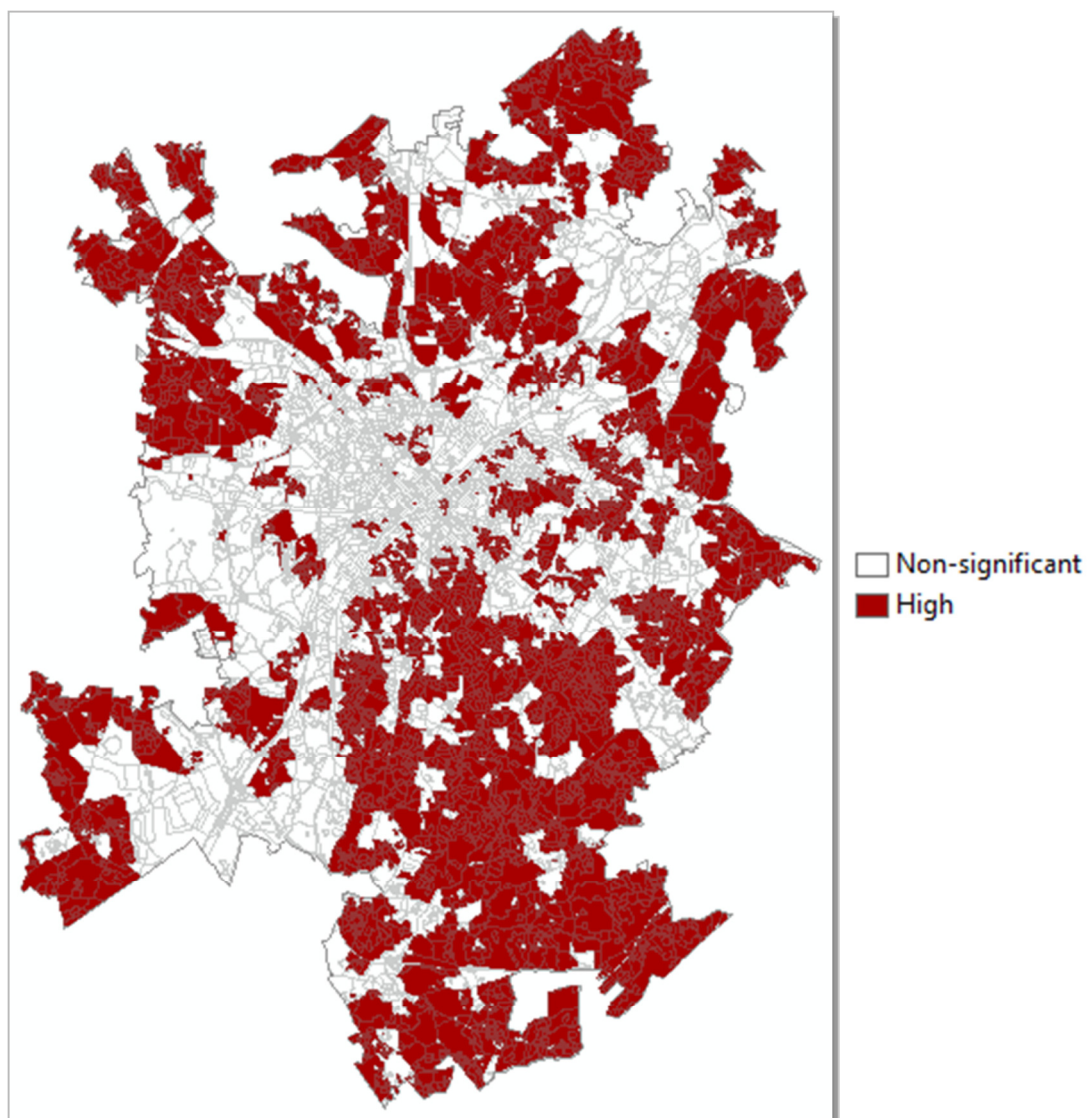


Figure 46: AMOEBA-based cluster for percentage of owner occupied houses using Census 2010 BLK data with Rook neighborhood definition

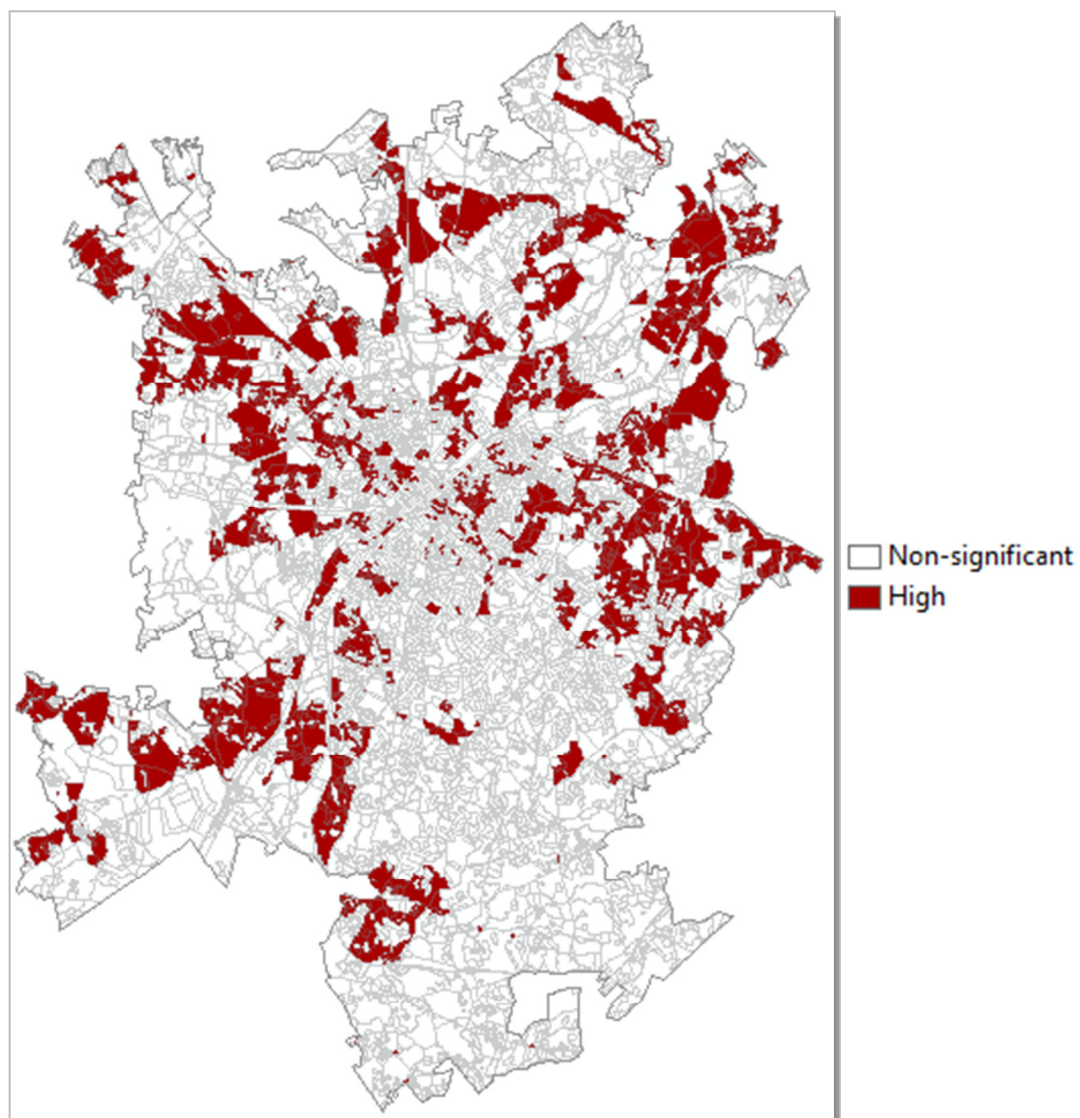


Figure 47: AMOEBA-based clusters for percentage of males aged 18-24 using Census 2010 BLK data with Rook neighborhood definition

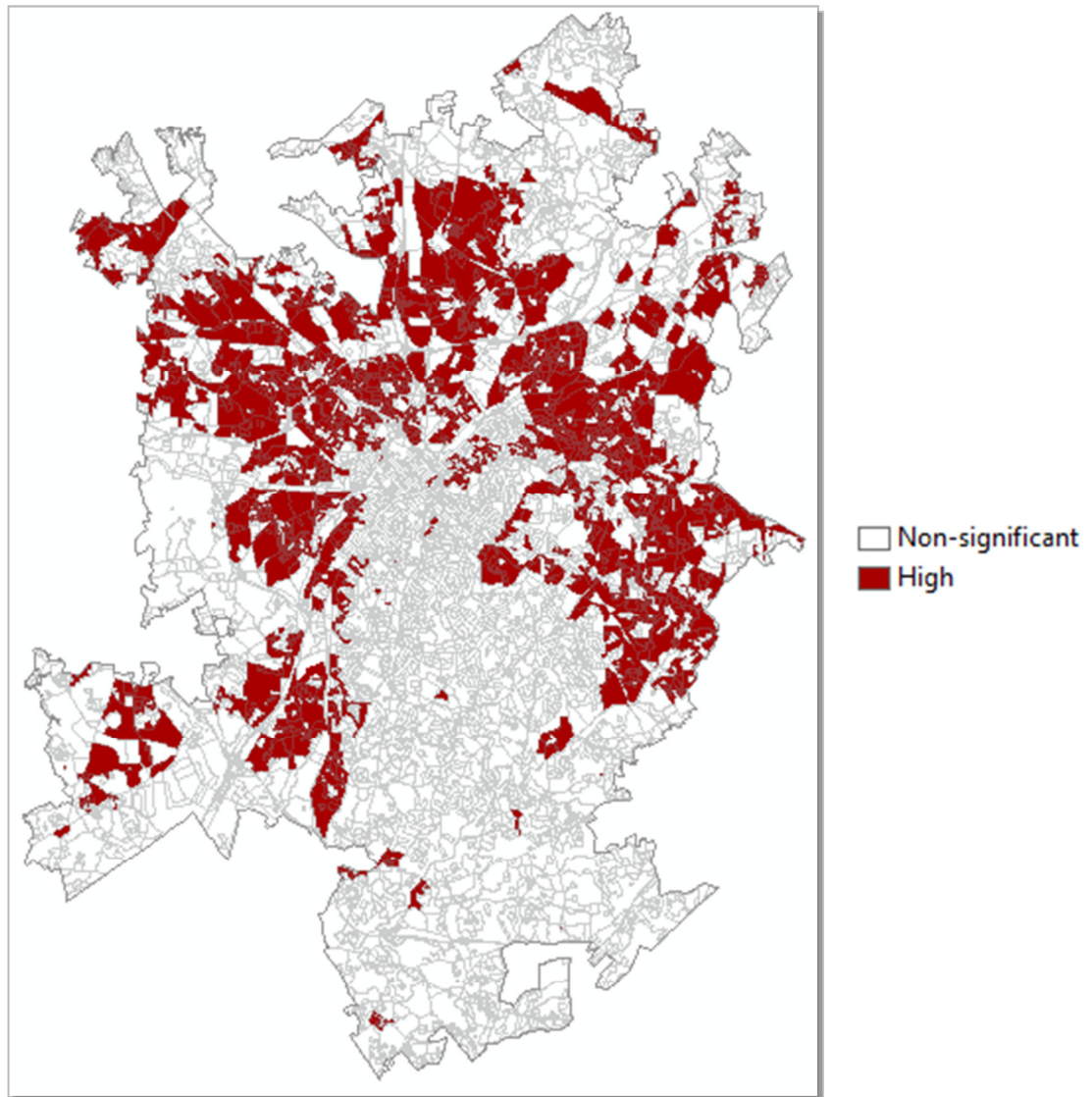


Figure 48: AMOEBA-based clusters for percentage of single-parent families using Census 2010 BLK data with Rook neighborhood definition

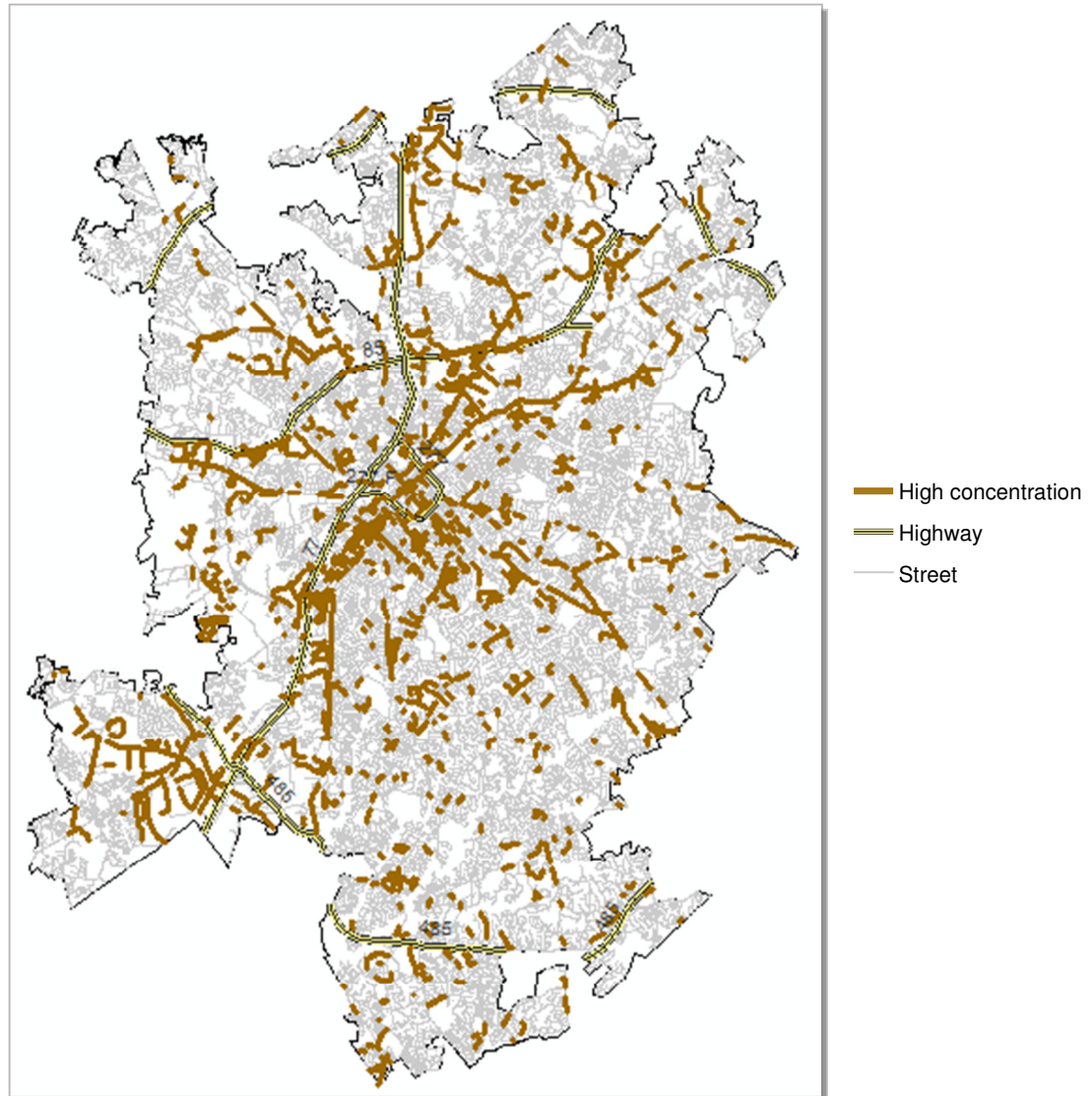


Figure 49: AMOEBA-based hotspots for business activities using 2008 business data

In this particular case study, it is observed that only hot spots (no cold spots) are detected by AMOEBA for some variables, including CAT, MVT, TFM, ethnic heterogeneity index, percentage of African-American population, percentage of owner occupied houses, percentage of males aged 18-24, percentage of single-parent families, and business activity. To better evaluate this situation, a comparison between AMOEBA and GeoDa on cluster detection using the same dataset of African-American population

was carried out as a test and the results are reported in Figure 50. The GeoDa cluster algorithm was set up using Rook neighborhood definition and 99 permutations. AMOEBA was also set up with 99 permutations test on the final set of non-overlap clusters with highest absolute G values. Details regarding spatial clustering algorithms for AMOEBA and GeoDa are summarized in Table 7. For the same data set of percentage of American-African population, AMOEBA detects only hot spots. GeoDa also detect only hot spots using the normal p value to test for significant. However, with a pseudo p value which serves to calibrate for the change in significance using different permutation values, both hot and cold spots are detected. While AMOEBA also uses pseudo p value, there exists a difference between AMOEBA pseudo p and GeoDa pseudo p (refer to Table 7 for details). The nature of AMOEBA which allows only hot spot detection in some cases, however, needs further careful examination to avoid any overstatement. The question that remains for future work is why AMOEBA is not detecting cold spots very readily for some variables in this case study and if it is due to an algorithmic knot, could AMOEBA be modified or enhanced to work with both hot and cold spots in all cases. In many practical problems of spatial analysis and modeling, spatial patterns of cold spots play an important role and could contribute interesting perspectives in studying associations governing spatial processes. For instance, in this particular study, spatial associations to low crime in contrast to those associated to high crime might offer potential insights into crime spatial patterns and thus criminal behaviors. The issue of AMOEBA performance will be a top priority in the future research agenda.

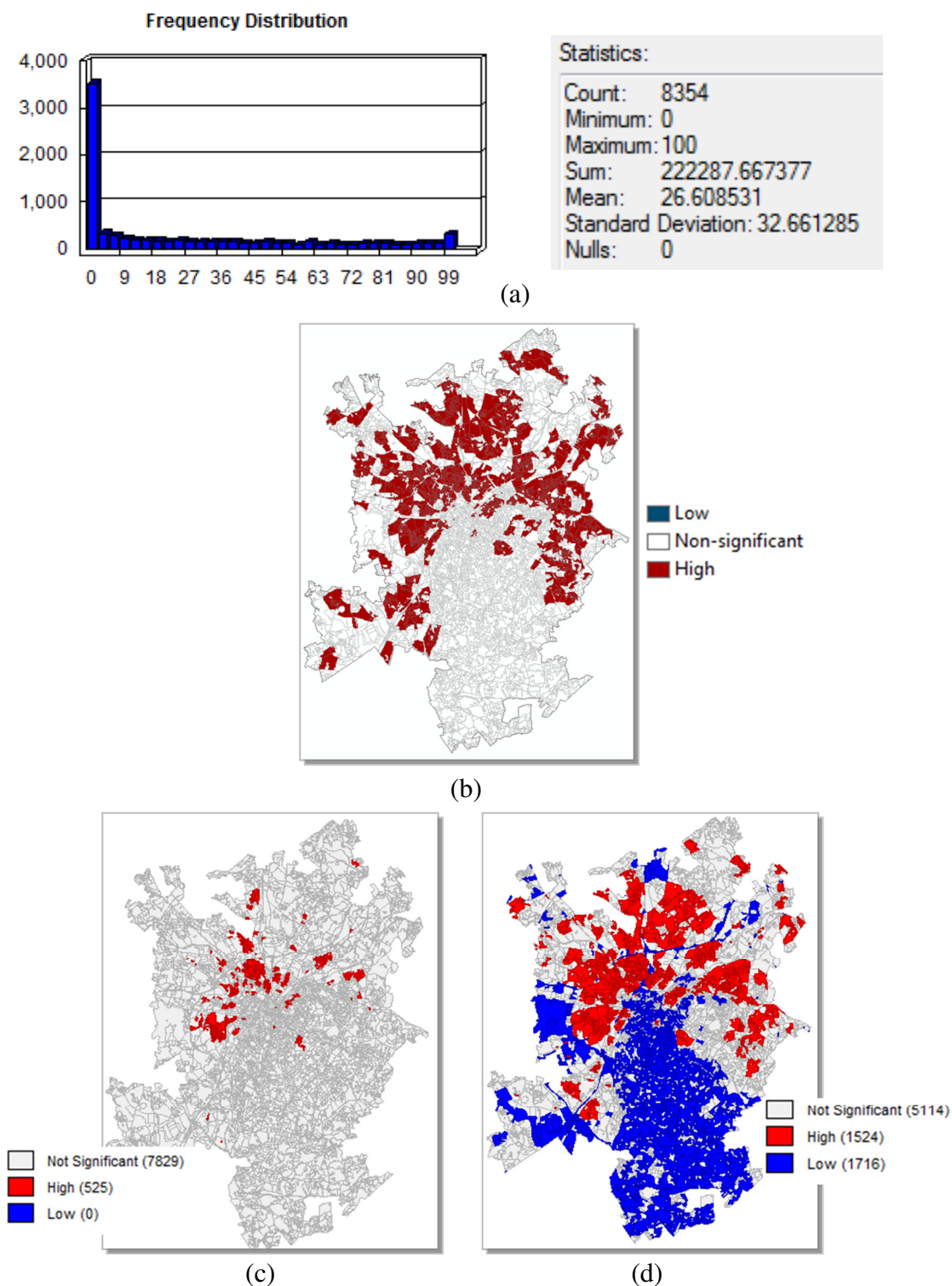


Figure 50: Percentage of AA population: (a) data statistics; (b) AMOEBA clustering result; (c, d) GeoDa Gi* clustering results using normal p value and pseudo p value respectively

Table 7: Summary of clustering algorithms for AMOEBA and GeoDa

AMOEBA	GeoDa
<p>Measures of spatial autocorrelation based on multi-directional search for neighbors contributing to the G^* value:</p> <ol style="list-style-type: none"> 1) For each cell i, identify its ecotope: <ol style="list-style-type: none"> a) Compute G_i^* b) Identify neighbors of i, estimate G_i^* for every combinations of cell i and neighbors, include the neighbor(s) if $\text{abs}(G_i^*)$ increase. Stop search when inclusion of neighbor doesn't increase G_i^*. 2) After ecotopes for each and every cell within the study area is identified, AMOEBA algorithm is continued by keeping the non-overlapping ecotopes with the highest G_i^* values. The exact probability that each ecotope has arisen by chance is then evaluated using a Monte Carlo-type permutation test. A large number of random permutations of the data set are generated. For each of these permutations, the G_i^* statistic is calculated for the ecotope. The p value is then calculated as the rank of the observed data set divided by the number of Monte Carlo realizations plus one. Only those ecotopes with p values below some predesignated level of significance are considered as true clusters. 	<p>Measures of spatial autocorrelation based on predetermined neighborhood definition contributing to the G^* value:</p> <ol style="list-style-type: none"> 1) Compute G_i for all i. 2) For each i, perform Monte Carlo N permutation test. Meaning compute N permuted G_i by drawing random samples of neighbors without including i 3) Keep track of every time a permuted G_i is greater than the observed G_i and call this #larger 4) If $N/2 > \text{\#larger}$: $p\text{-value} = (1 + \text{\#larger}) / (1 + N)$ Otherwise $p\text{-value} = (1 + N - \text{\#larger}) / (1 + N)$ <p>Only those with p-value below some predesignated level of significance are considered as clusters.</p>

6.3.2 Modeling Spillover Effects of Spatial Dependence Structures

Previous discussions on the SpatialARMED framework in Chapter 5, particularly Section 5.2.2, have stressed the necessity to model spatial spillover effects of spatial dependence structures, represented by hot and cold spots, as well as of crime generators and attractors in order to generate predicates expressing the indirect spatial relations of crime to its potential associated factors. For the hot and cold spot spatial spillover effects, the model in Equations (3) and (4), which is estimated as the sum of inverse distance weighted individual G^* for all cluster members, is applied for this case study. For crime generator and attractor spillover effects, the model in Equation (5) is used. For this particular case study, attributes of these crime generators and attractors such as size and popularity, are not included. Their spillover effects are purely a function of inverse distances. For all cases, the spillover effects are estimated as functions of the inversed street network distance raised to a power of 1 (i.e. $\alpha = 1$ in Equations 3, 4, and 5) and at the level of street blocks.

The spillover impacts of dangerous streets due to crime of all types, motor vehicle thefts, and thefts from motor vehicle are shown in Figure 51 to Figure 53 , respectively. On the left-hand side of these figures are AMOEBA-based clusters of high crime activity of all types, of MVT, and of TFM, respectively. Examination of individual G^* values for members of the clusters, which are highlighted with yellow markers in these figures, helps to visualize the internal structure of each cluster. On the right-hand side of these figures are maps of standardized spatial spillover effects. These indeed portray areas under strong spatial spillover effects of criminal activities, or in other words, the core of criminal spatial dependence structures.

In a similar mapping schema, Figure 54 to Figure 64 present the results of spatial spillover effects for all associate variables listed in Table 6. Those of crime generators and attractors are presented in Figure 65 to Figure 69.

By looking at these spatial spillover maps, the argument could be made on the legitimacy of the spatial spillover effects being modelled one way or the other. Please refer to Section 5.2.2. for discussions on different modeling approaches, which very much depend on the nature of the spillover phenomenon itself, and the need for a sensitivity test. In this analysis, a particular model using Equations 3, 4, and 5 with $\alpha = 1$ for all variables is used, to serve as a demonstration for SpatialARMED implementation. In more practical cases of SpatialARMED, it is suggested to rely on sensitivity testing for the most optimal model of spatial spillover effect applicable to each variable under study.

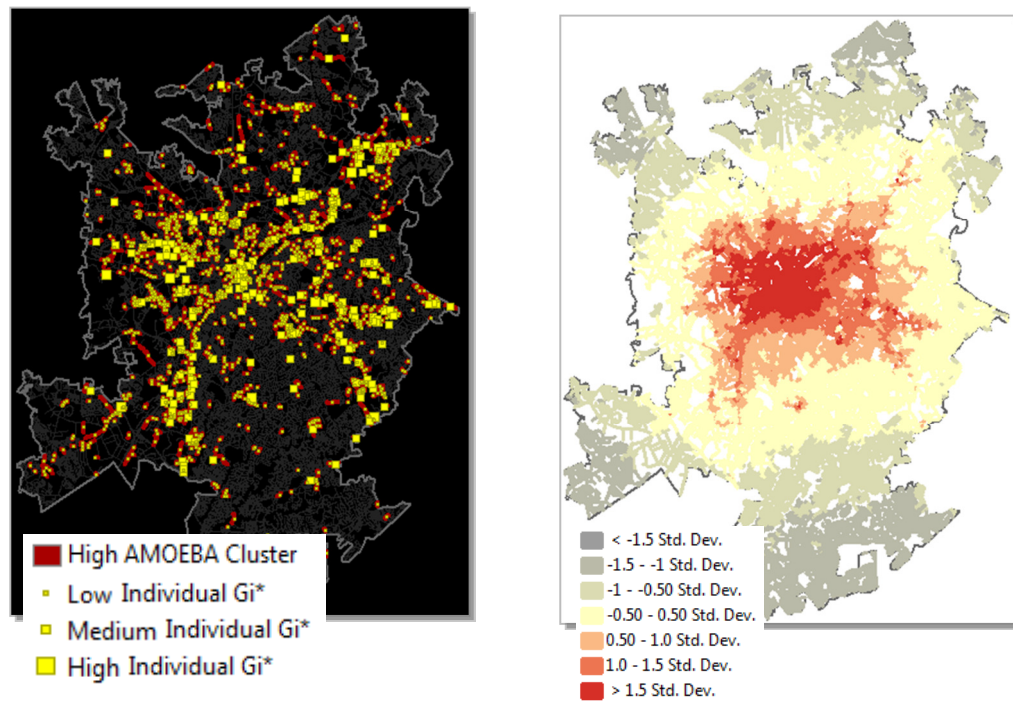


Figure 51: AMOEBA-based detected streets of high crime (left) and spatial spillover effect (right)

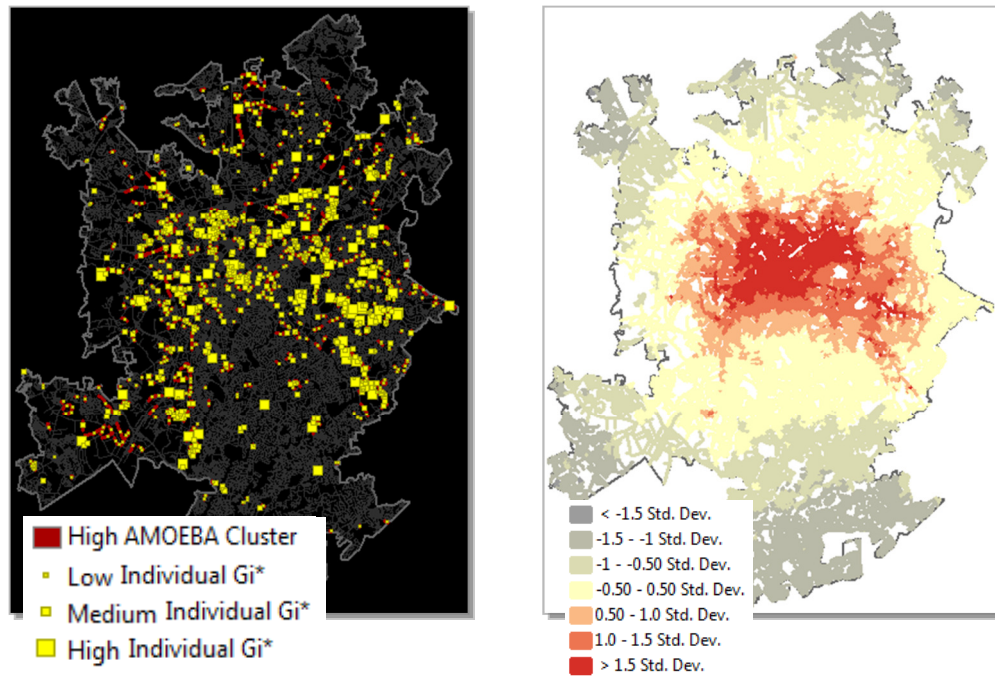


Figure 52: AMOEBA-based detected streets of high motor vehicle thefts (left) and spatial spillover effect (right)

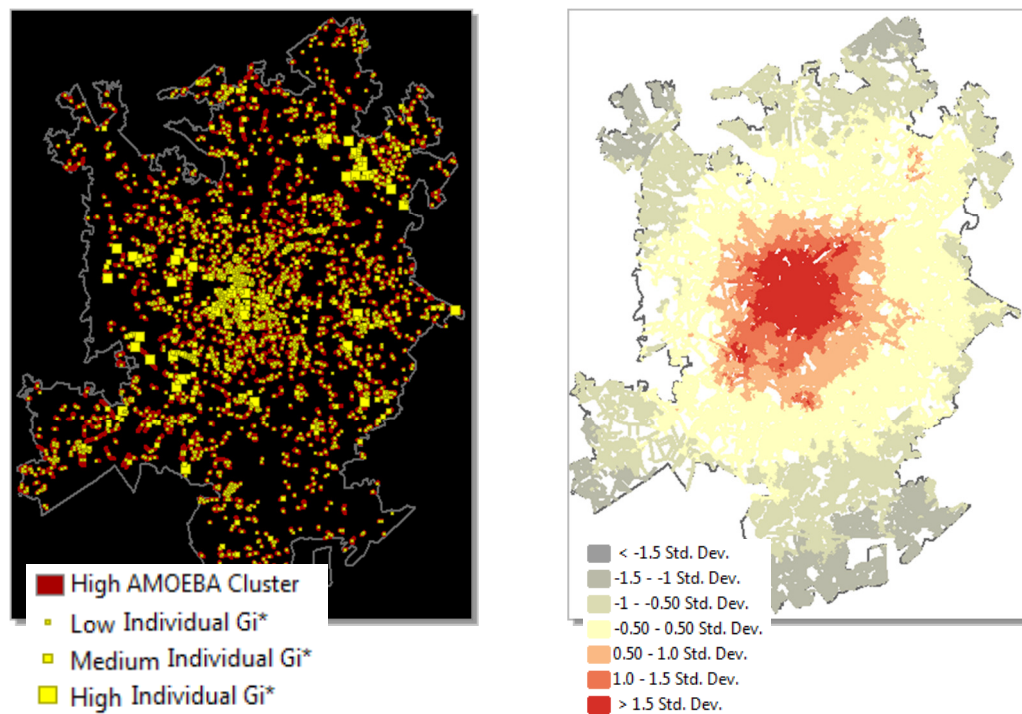


Figure 53: AMOEBA-based detected streets of high thefts from motor vehicle (left) and spatial spillover effect (right)

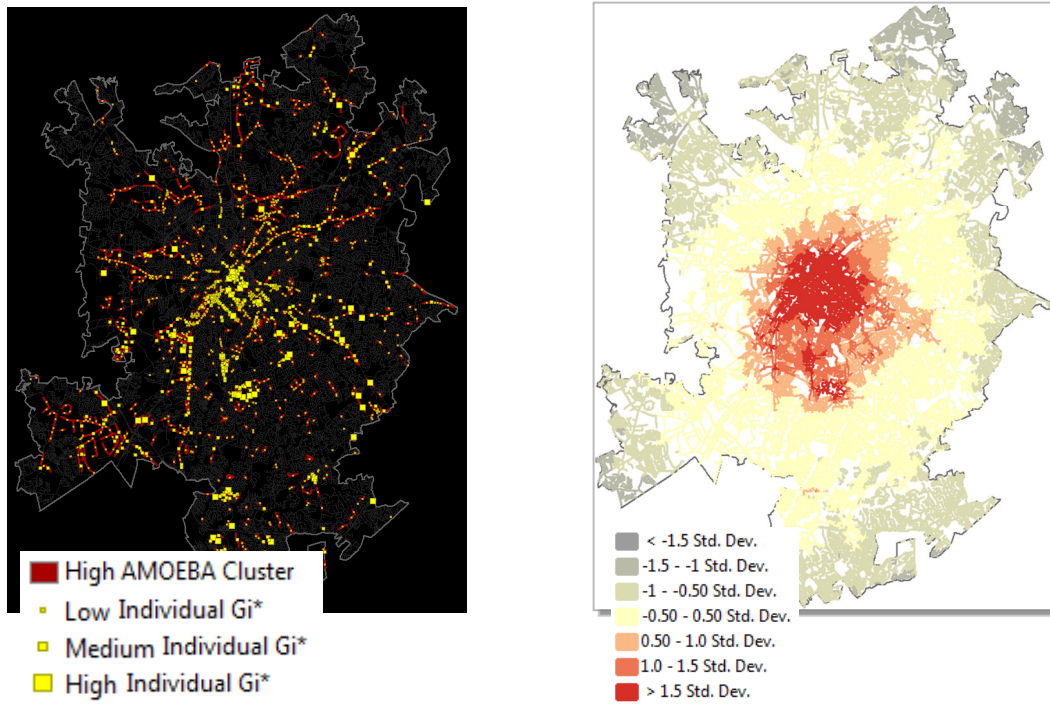


Figure 54: AMOEBA-based detected streets of high commercial activity (left) and spatial spillover effect (right)

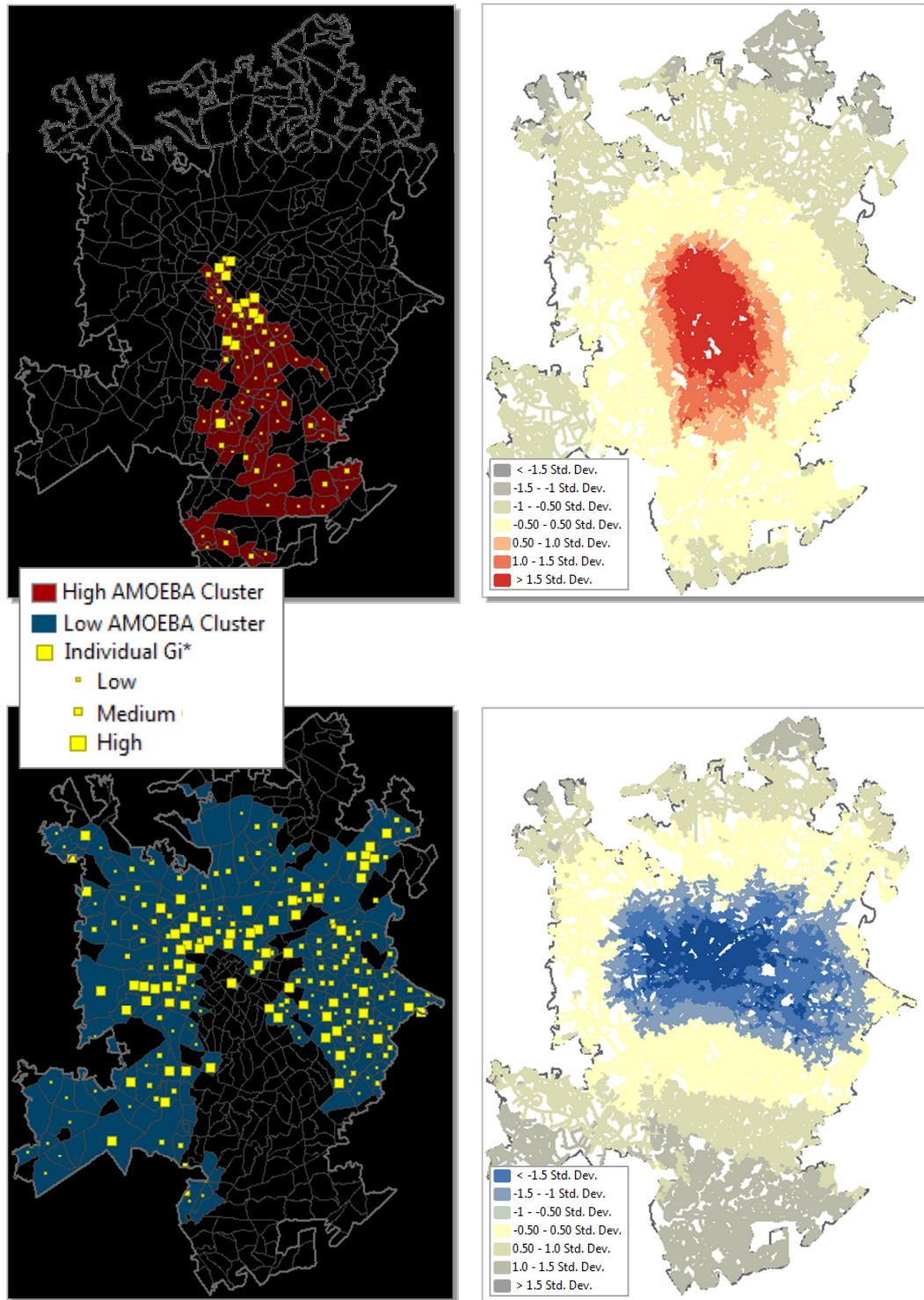


Figure 55: Per Capital Income: AMOEBA-based detected clusters of high values (upper left) with its spatial spillover effect (upper right) and AMOEBA-based detected clusters of low values (bottom left) with its spatial spillover effect (bottom right)

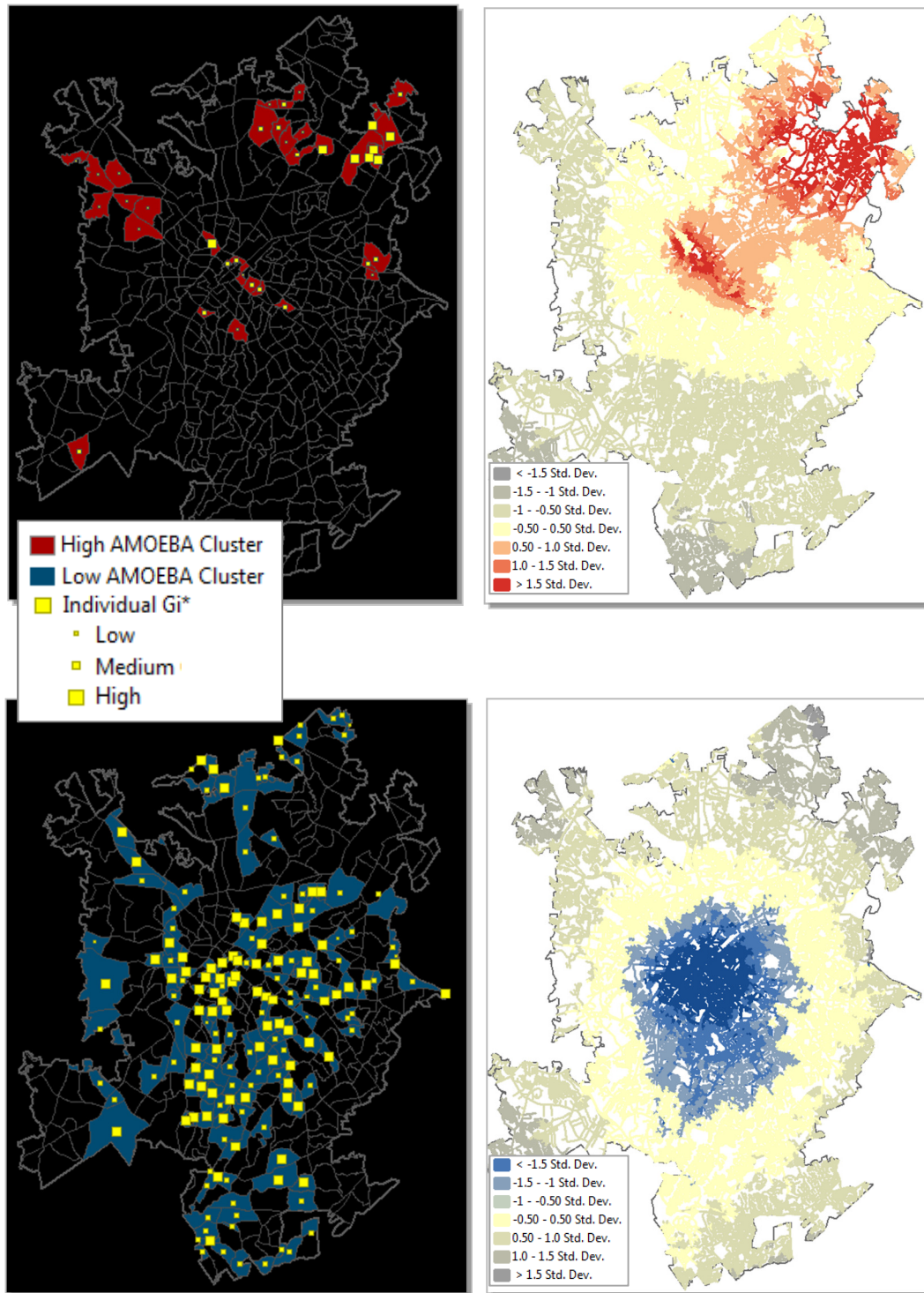


Figure 56: Percentage of population with high school degree or above: AMOEBA-based detected clusters of high values (upper left) with its spatial spillover effect (upper right) and AMOEBA-based detected clusters of low values (bottom left) with its spatial spillover effect (bottom right)

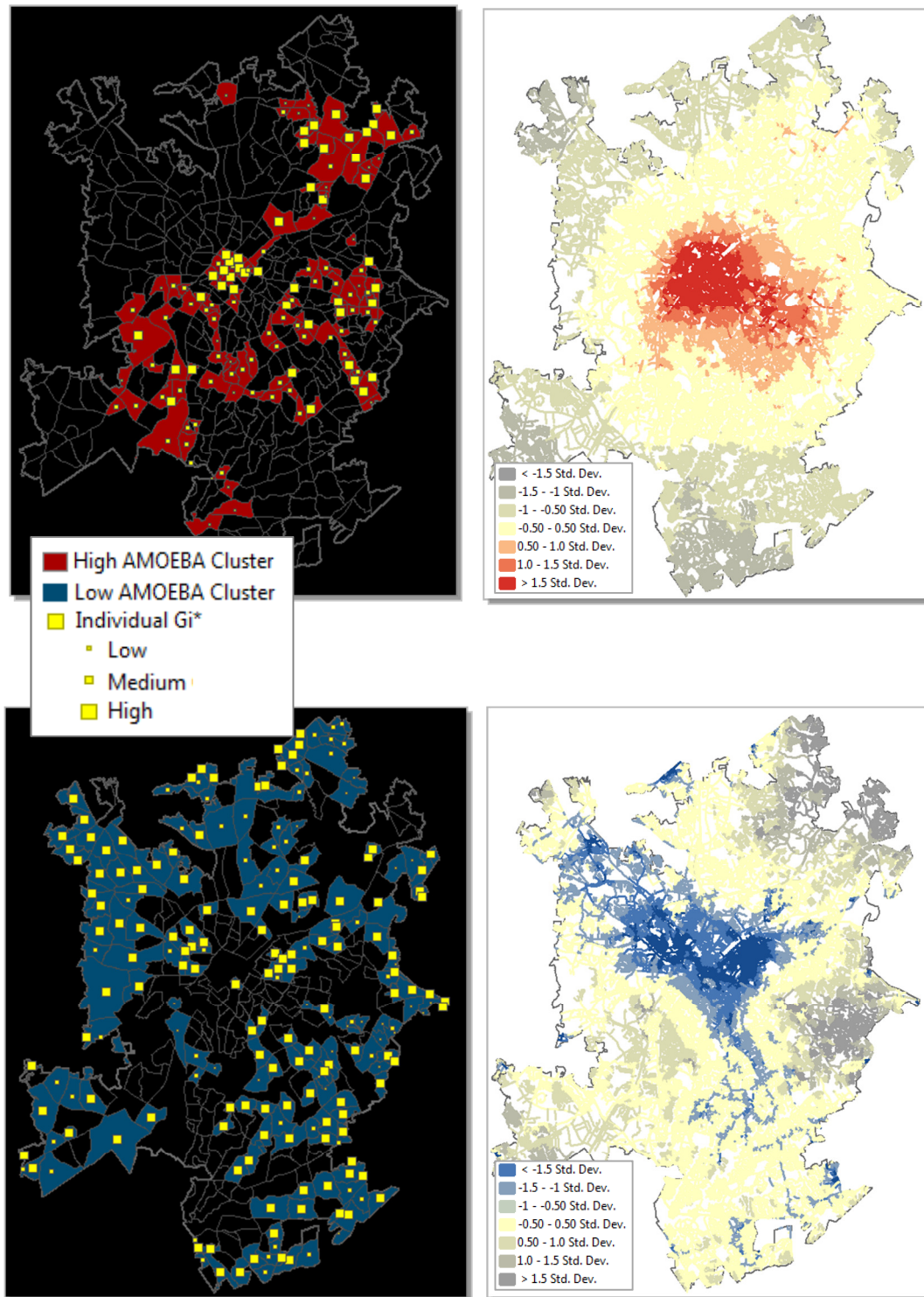


Figure 57: Percentage of housing structures of 3 or more units: AMOEBA-based detected clusters of high values (upper left) with its spatial spillover effect (upper right) and AMOEBA-based detected clusters of low values (bottom left) with its spatial spillover effect (bottom right)

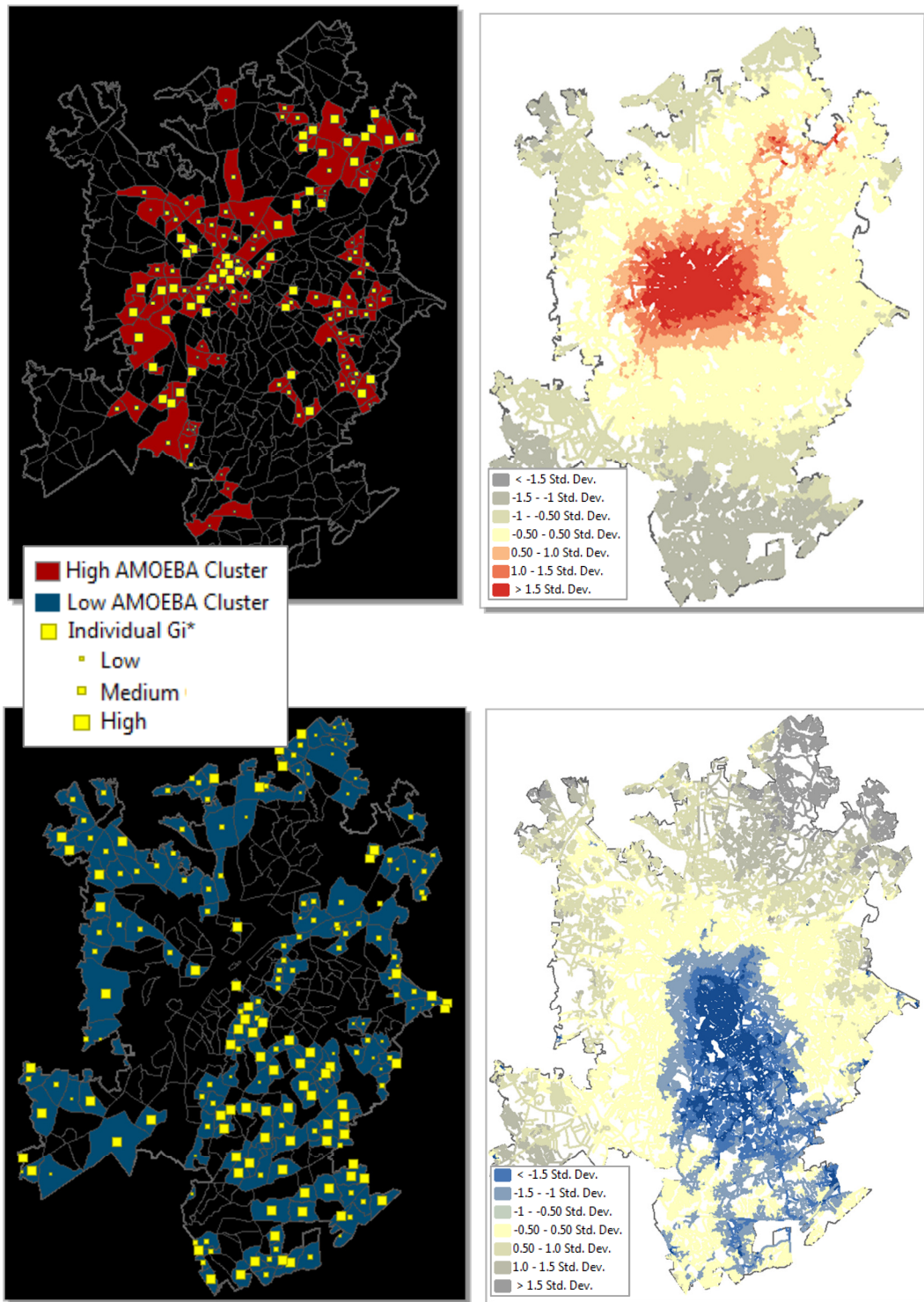


Figure 58: Percentage of population who rent and move in less than 5 year: AMOEBA-based detected clusters of high values (upper left) with its spatial spillover effect (upper right) and AMOEBA-based detected clusters of low values (bottom left) with its spatial spillover effect (bottom right)

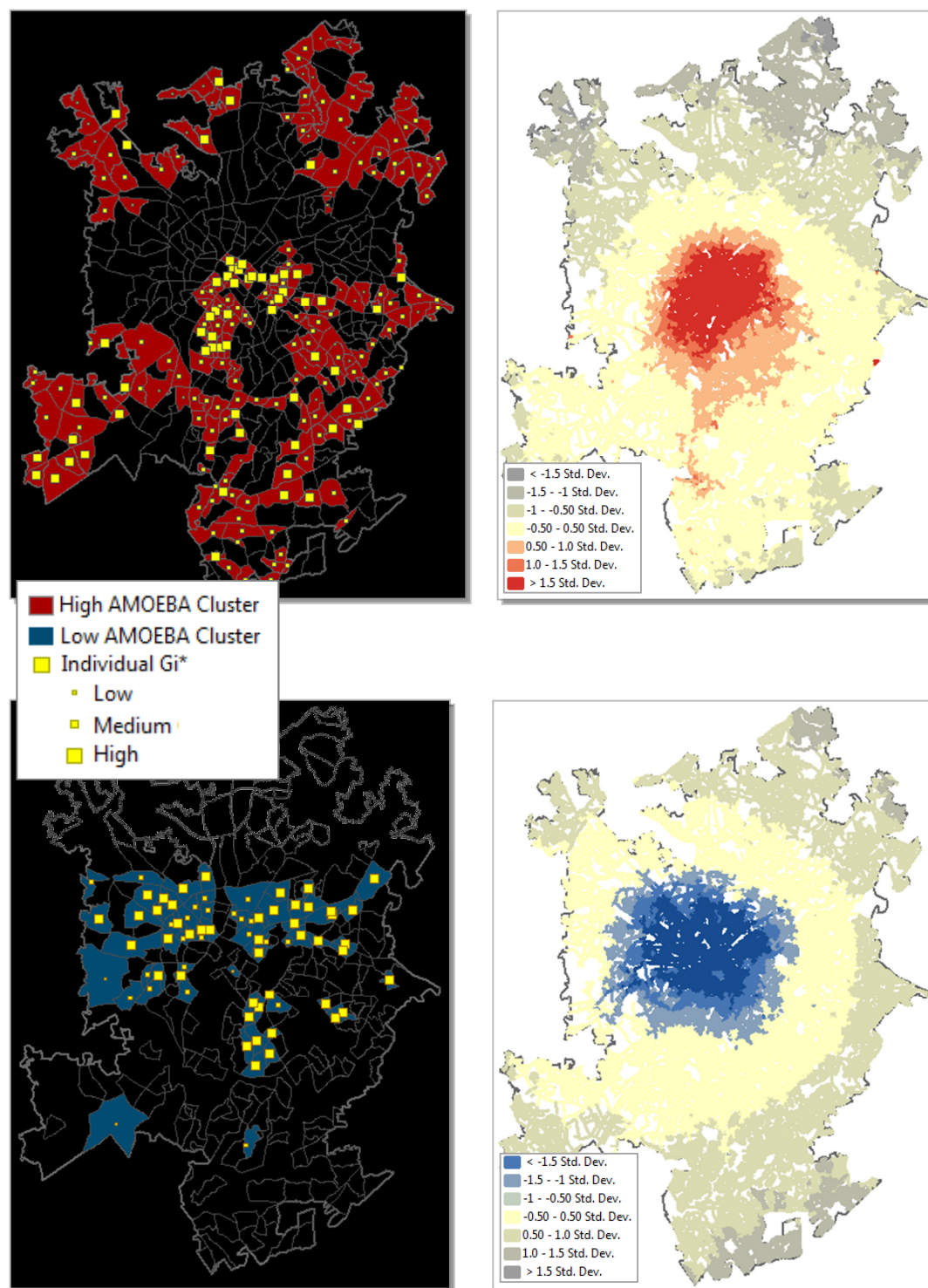


Figure 59: Percentage of population employed: AMOEBA-based detected clusters of high values (upper left) with its spatial spillover effect (upper right) and AMOEBA-based detected clusters of low values (bottom left) with its spatial spillover effect (bottom right)

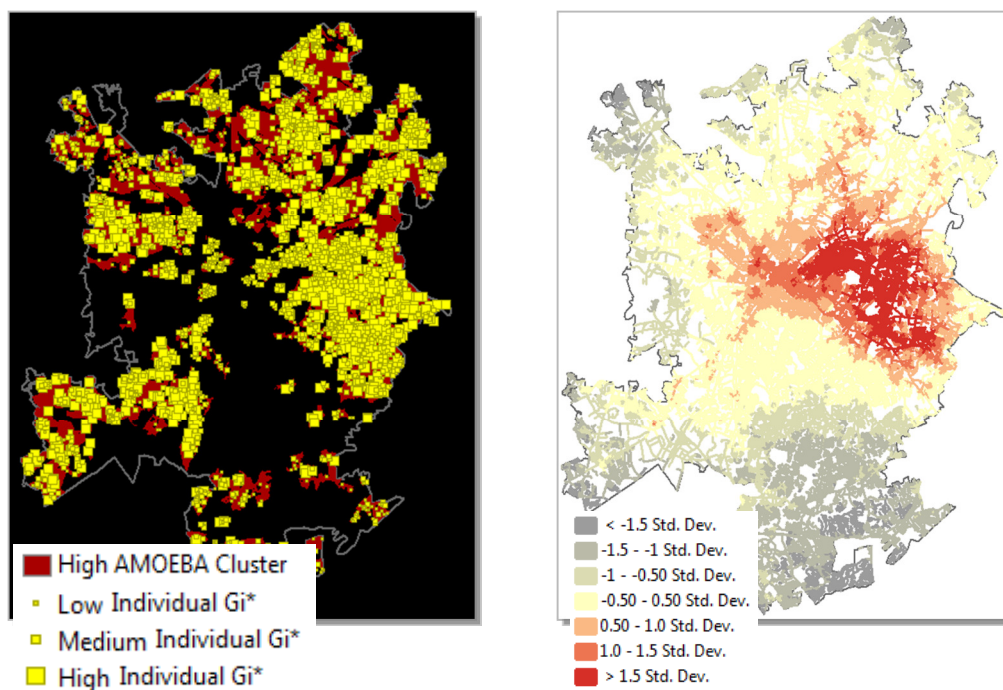


Figure 60: Heterogeneity Index: AMOEBA-based detected clusters of high values (left) and spatial spillover effect (right)

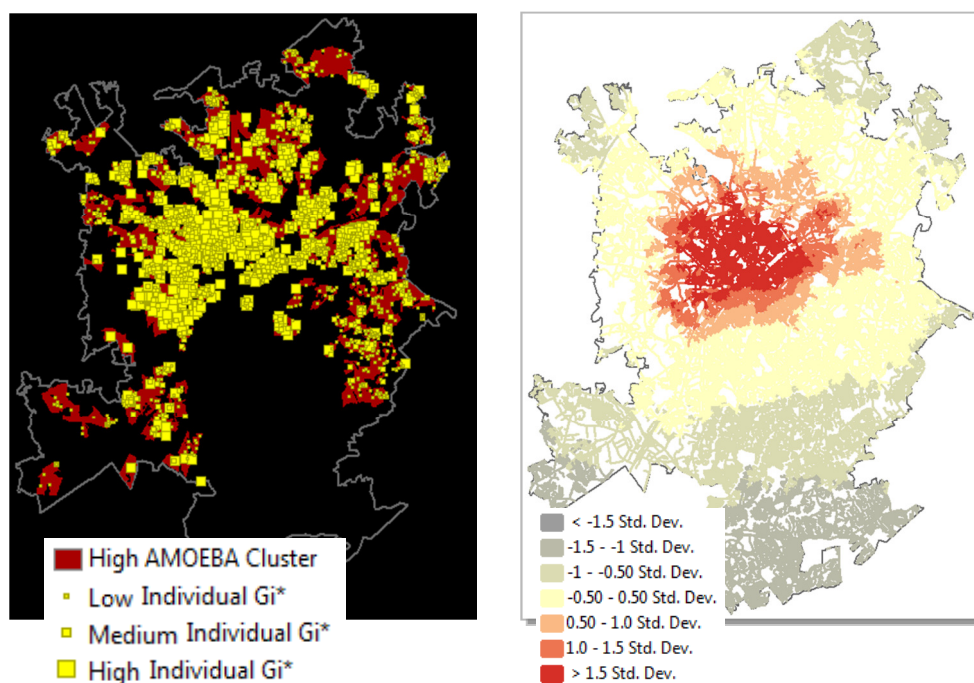


Figure 61: Percentage of African-American population: AMOEBA-based detected clusters of high values (left) and spatial spillover effect (right)

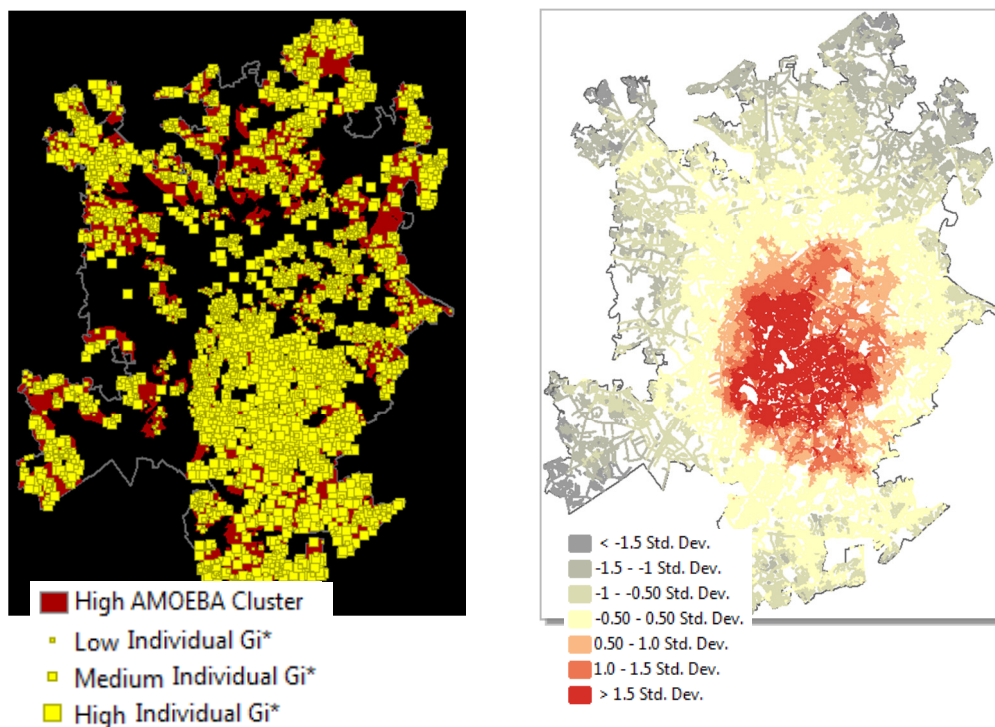


Figure 62: Percentage of owner occupied homes: AMOEBA-based detected clusters of high values (left) and spatial spillover effect (right)

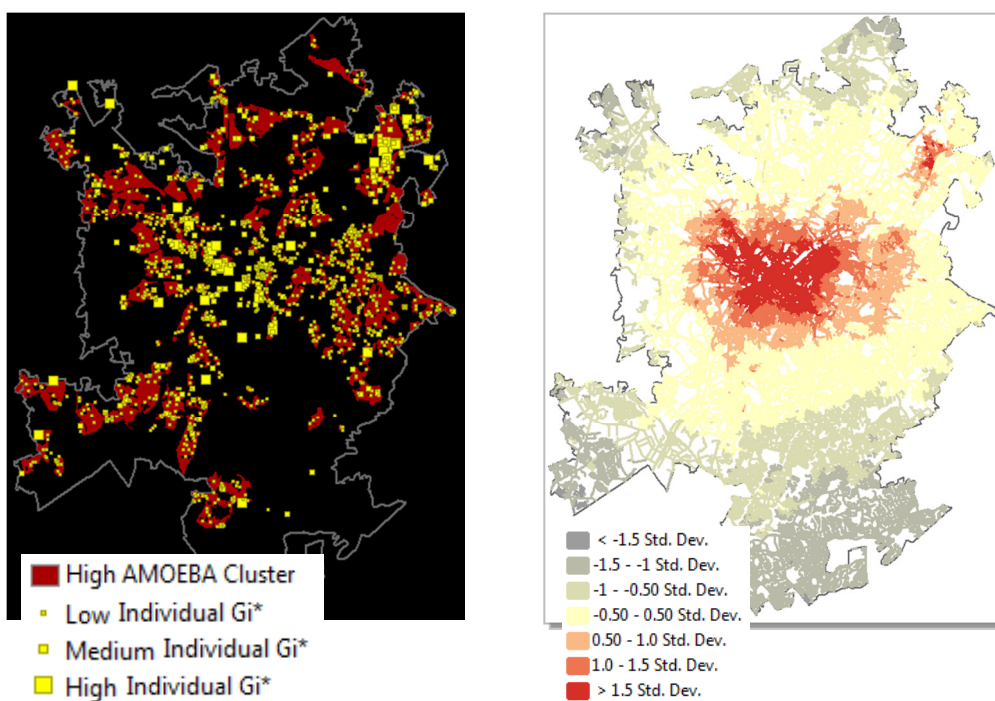


Figure 63: Percentage of population who are males aged 18-24: AMOEBA-based detected clusters of high values (left) and spatial spillover effect (right)

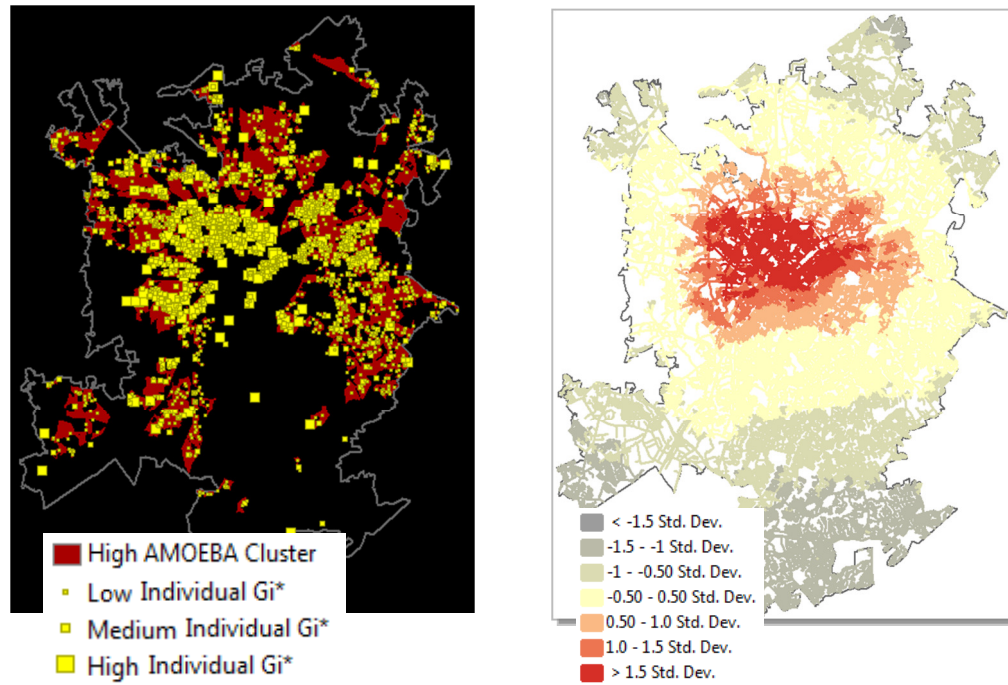


Figure 64: AMOEBA-based detected cluster of high percentage of single-parent families (left) and spatial spillover effect (right)

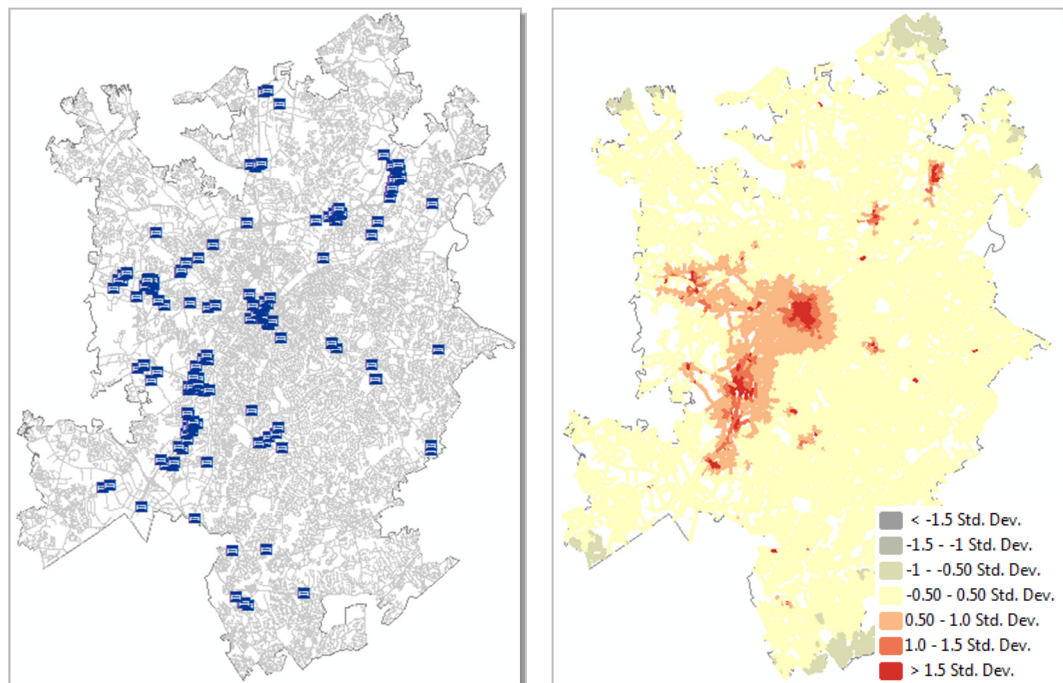


Figure 65: Location of hotels and models (left) and spatial spillover effect (right)

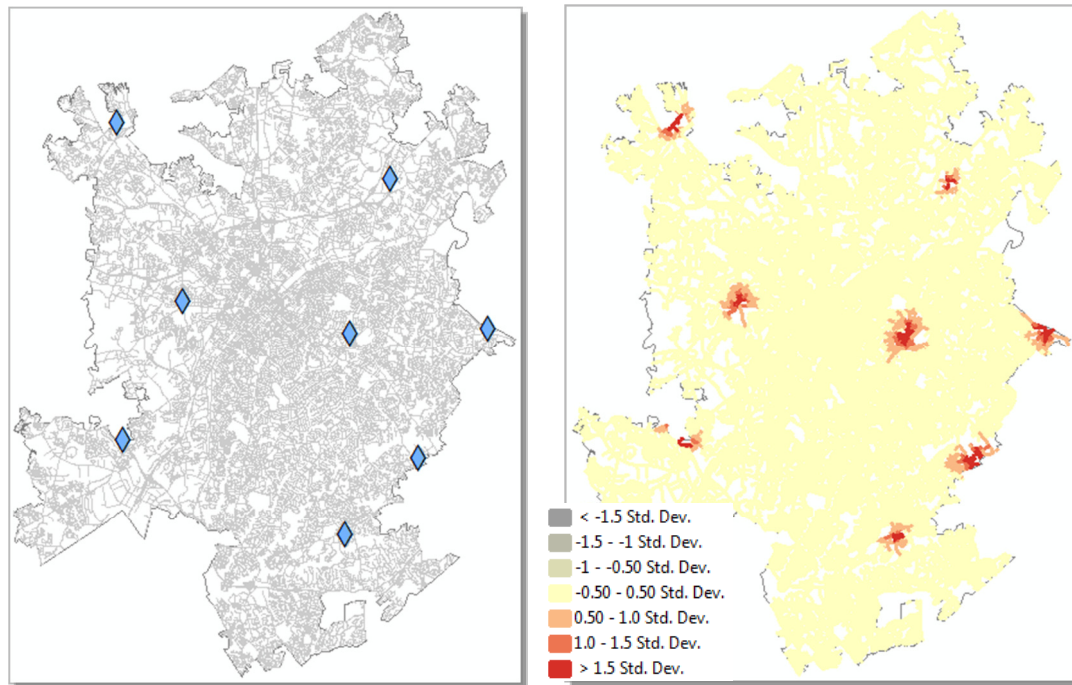


Figure 66: Location of Wal-Mart super centers (left) and spatial spillover effect (right)

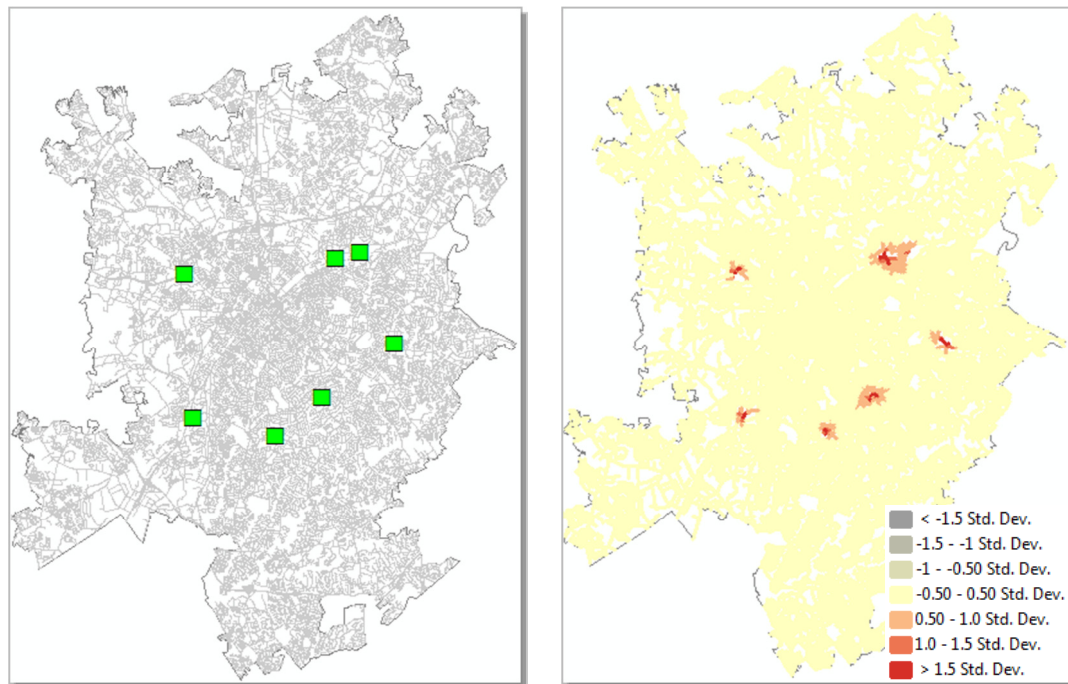


Figure 67: Location of shopping malls (left) and spatial spillover effect (right)

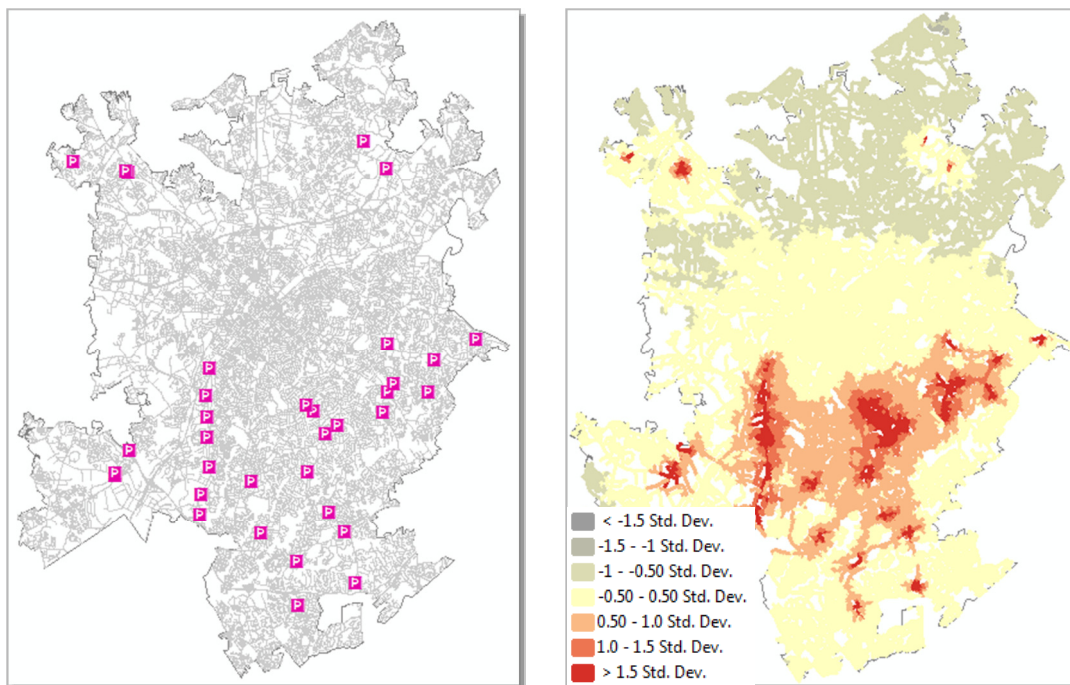


Figure 68: Location of park-and-ride facilities (left) and spatial spillover effects (right)

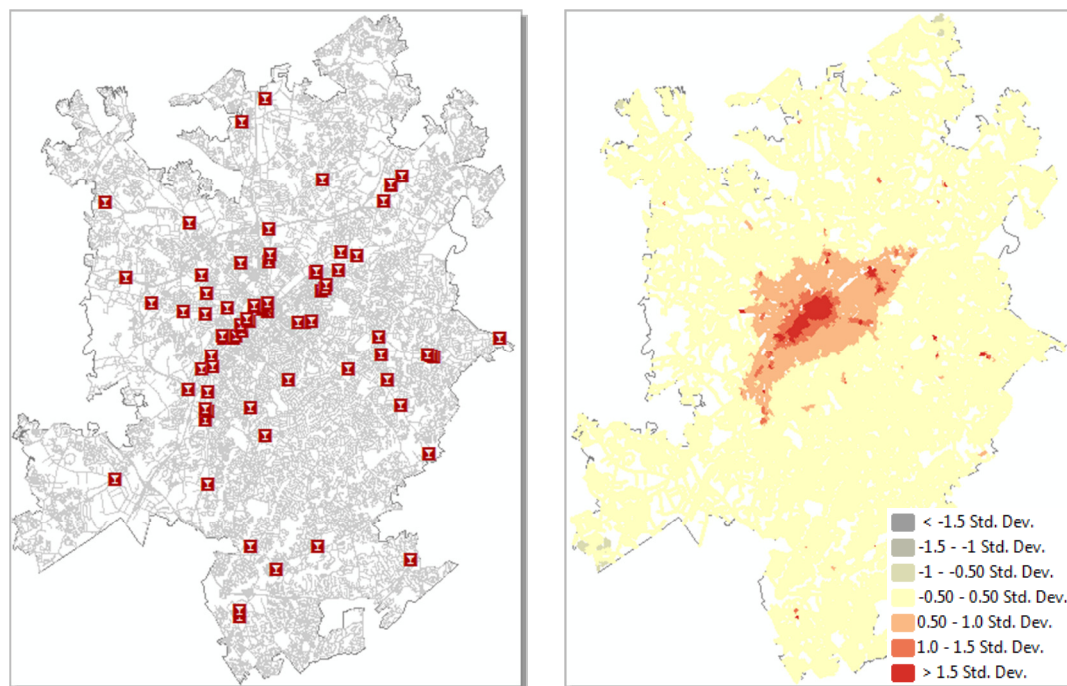


Figure 69: Location of Alcohol Drinking Places (left) and spatial spillover effects (right)

6.4 SpatialARMED Level 3: Predication

Up to this point, the spatial dependence structures of crime and associate variables have been quantified. In addition, the spatial spillover effects of hot and cold spots with respect to these variables have been modelled. The next step in the SpatialARMED process involves generating predicates and the final table to mine.

6.4.1 Define Unit of Tuples and Generate Predicates with Numeric Values

In order to carry out spatial join operations on the identified spatial dependence structures and spillover effects, unit of tuples should be properly defined as discussed in Section 5.3.1. The issue of MAUP is particularly relevant to this case study because data are collected at various resolutions, including block group, block, and street block. Tuples in this case should be the lowest resolution unit, which is the block group, in order to avoid the biased estimation of support and confidence of frequent item sets and rules.

Once the unit of tuples is decided (i.e. block groups), associate attributes of any higher spatial resolution units (i.e. streets and blocks) need to be mapped onto them in order to generate predicates for mining. In order to achieve this, various additional predicates for block groups are generated to indicate percentage of the area or percentage of its total street length overlapping with the identified spatial dependence structure (SS) and spillover effects (SIM) for variables of blocks and street blocks. Table 8 provides details regarding this attribute manipulation process. As a result, spatial predicates and their numeric values for block groups are generated regarding characteristics of crime of all types (CAT), motor vehicle thefts (MVT), and thefts from motor vehicle (TFM), business activity, and other socio-economic and demographic variables. To make it easier to understand, the predicates are organized into three different types as shown in Table 9.

Type 1 predicates are those associated with block group variables (i.e. the one of lowest spatial resolution within this study case) and present block-group attributes, namely income, education, employment, MU housing, and rental population who move in less than 5 years. They take the G^* values of the corresponding block groups as the result of AMOEBA clustering detection on these variables. Type 2 predicates are for variables transferred from street blocks and census blocks (i.e. the ones with higher spatial resolution than block groups). Type 2 predicate values for each block group as the mining unit will be the percentage of the block group's area or street length overlapping with the variable's spatial dependence structures. Type 3 predicates indicate spatial spillover effects (SIM). In the subsequent steps of predication, all these predicates representing semantic and spatial attributes for the mining unit (i.e. block groups in this case) will be mapped to linguistic expressions and present nominal measures on the scale from high to low.

Table 8: Process of generating predicates with block group as unit of tuples for the dangerous street mining task (SS stands for Spatial Dependence Structure)

Unit of Tuple	Spatial operation: overlap with:	Predicate (name, numeric value)
Block group	SS (CAT)	(CAT, % of total BLG street length has $G^*_{CAT} > 0$)
	Spillover of high CAT	(High CAT impact, Average SIM_{P-CAT} for all BLG streets)
	SS (MVT)	(MVT, % of total BLG street length has $G^*_{MVT} > 0$)
	Spillover of high MVT	(High MVT impact, Average SIM_{P-MVT} for all streets within BLG)
	SS (TFM)	(TFM, % of total BLG street length has $G^*_{TFM} > 0$)
	Spillover of high TFM	(High TFM impact, Average SIM_{P-TFM} for all streets within BLG)

SS (Business location)	(Business activity, % of total BLG street length has AMOEBA $G^*_{\text{Business}} > 0$)
Spillover of high concentration of business	(High business impact, Average $\text{SIM}_{\text{P-Business}}$ for all streets within BLG)
SS (Per-Capital Income)	(Income, AMOEBA G^*_{Income})
Spillover of high income	(High Income impact, Average $\text{SIM}_{\text{P-Income}}$ for all streets within BLG)
Spillover of low income	(Low income impact, Average $\text{SIM}_{\text{N-Income}}$ for all streets within BLG)
SS (% of population has high school degree or higher)	(Education, $G^*_{\text{Education}}$)
Spillover effect of high education	(High education impact, Average $\text{SIM}_{\text{P-Education}}$ for all streets within BLG)
Spillover effect of low education	(Low education impact, Average $\text{SIM}_{\text{N-Education}}$ for all streets within BLG)
SS (% of population employed)	(Employment, AMOEBA $G^*_{\text{Employment}}$)
Spillover effect of high employment	(High employment impact, Average $\text{SIM}_{\text{P-Employment}}$ for all BLG streets)
Spillover effect of low employment	(Low employment impact, Average $\text{SIM}_{\text{N-Employment}}$ for all BLG streets)
SS (% of housing unit has multiple (>3) unit structure)	(MUHousing, AMOEBA $G^*_{\text{MUHousing}}$)
Spillover effect of high MU housing	(High MUHousing impact, Average $\text{SIM}_{\text{P-MUHousing}}$ for all BLG streets)
Spillover effect of low MU housing	(Low MUHousing impact, Average $\text{SIM}_{\text{N-MUHousing}}$ for all BLG streets)
SS (% of population who rent and move in <5 years)	(Unstable Rent, AMOEBA $G^*_{\text{Unstable Rent}}$)
Spillover of high unstable rent	(High unstable rent impact, Average $\text{SIM}_{\text{P-Unstable Rent}}$ for all BLG streets)
Spillover of low unstable rent	(Low, unstable rent impact, Average $\text{SIM}_{\text{N-Unstable Rent}}$ for all streets within BLG)
SS (Population density)	(Population density, % of BLG area has AMOEBA $G^*_{\text{PopDen.}} > 0$)
Spill over effect of high density	(High population density impact, Average $\text{SIM}_{\text{P-PopDen}}$ for all streets within BLG)

SS (% of African American population)			(African American, % of BLG area has AMOEBA $G^*_{AA} > 0$)
Spillover of concentration of AA	high		(High AA impact, Average SIM_{P-AA} for all streets within BLG)
Spillover of concentration of AA	low		(Low AA impact, Average SIM_{N-AA} for all streets within BLG)
SS (% of male 17-28 population)			(Male 17-24, % of BLG area has AMOEBA $G^*_{Male1724} > 0$)
Spillover of concentration of male 17-24	high		(High male 17-24, Average $SIM_{P-Male1724}$ for all streets within BLG)
Spillover of concentration of male 17-24	low		(Low male 17-24, Average $SIM_{N-Male1724}$ for all streets within BLG)
SS (% of home owned)			(Home owned, % of BLG area has AMOEBA $G^*_{HOWN} > 0$)
Spillover impact of home owned	high		(High home owned, Average SIM_{P-HOWN} for all streets within BLG)
Spillover impact of home owned	low		(Low home owned, Average SIM_{N-HOWN} for all streets within BLG)
SS (% of single parent family)			(SParent family, % of BLG area has $G^*_{SF} > 0$)
Spillover impact of concentration of single parent family	high		(High SParent family, Average SIM_{P-SF} for all streets within BLG)
Spillover impact of concentration of single parent family	low		(Low SParent family, Average SIM_{N-SF} for all streets within BLG)
SS (Heterogeneity index)			(Heterogeneity, % of BLG area has AMOEBA $G^*_{HeteInx} > 0$)
Spillover of heterogeneity	high		(High heterogeneity, Average $SIM_{P-HeteInx}$ for all streets within BLG)
Spillover of heterogeneity	low		(Low heterogeneity, Average $SIM_{N-HeteInx}$ for all streets within BLG)
Spillover of Wal-mart			(Wal-mart impact, Average $SIM_{POI-Walmart}$ for all streets within BLG)
Spillover of Mall			(Mall impact, Average $SIM_{POI-Mall}$ for all streets within BLG)
Spillover of drinking places	alcoholic		(Drinking place impact, Average $SIM_{POI-DnkPlacI}$ for all streets within BLG)

Spillover of hotels and motels	Hotel & Motel impact, Average $SIM_{POI-HMtel}$ for all streets within BLG)
Spillover of ParkNride facilities	ParkNRide impact, Average $SIM_{POI-ParkNRide}$ for all streets within BLG)

Table 9: Summary of predicate types for dangerous street SAR mining

Type	Predicate numeric value X	For predicates associated with:
1	$X = AMOEBA\ G^*$	Income, Education, Employment, MU Housing, and Unstable Rent
2	$X = \% \text{ of BLG area has } AMOEBA\ G^* > 0$ or $X = \% \text{ of street length has } AMOEBA\ G^* > 0$	African American, Heterogeneity, Single-parent family, HomeOwn, and male 17-24 Crime of all types, MVT, TFM, and Business activity
3	Based on average value of SIM_P , SIM_N , or SIM_{POI}	Spatial spillover impact result of the variables

6.4.2 Numeric-to-Nominal Mapping with Crisp Boundary

In order to facilitate the item set count and association rule mining, there is a need to map the numeric values to nominal categories, ranging from High to Low. This can be accomplished by classification using crisp boundaries or fuzzy boundaries with mapping mechanisms proposed in Section 5.3.3.

Table 10 describes the classification scheme applied in this case study. Figure 70 to Figure 72 depict standardized values histogram and crisp boundary mapping mechanism for Type 1, Type 2, and Type 3 predicates, respectively.

Table 10: Summary of numeric-to-nominal mapping mechanism using crisp boundary for all predicates

Type	Predicate numeric value X	Mapping Mechanism
1	X = AMOEBA G*	Number of classes = 3 BreakPoint = {0} H: $X > 0$ M: $X = 0$ L: $X < 0$
2	X = % of BLG area has AMOEBA G* > 0 or X = % of street length has AMOEBA G* > 0	Number of classes = 3 BreakPoint = {BP1, BP2} H: $X > BP2$ MH: $BP1 < X < BP2$ L: $X = 0$
3	X = $\overline{SIM_P}$, $\overline{SIM_N}$, or $\overline{SIM_{POI}}$	Number of classes = 3 BreakPoint = {BP1, BP2} H: $X > BP2$ M: $BP1 < X < BP2$ L: $X < BP1$

Type 1 predicate – BLG variables: X = AMOEBA G*

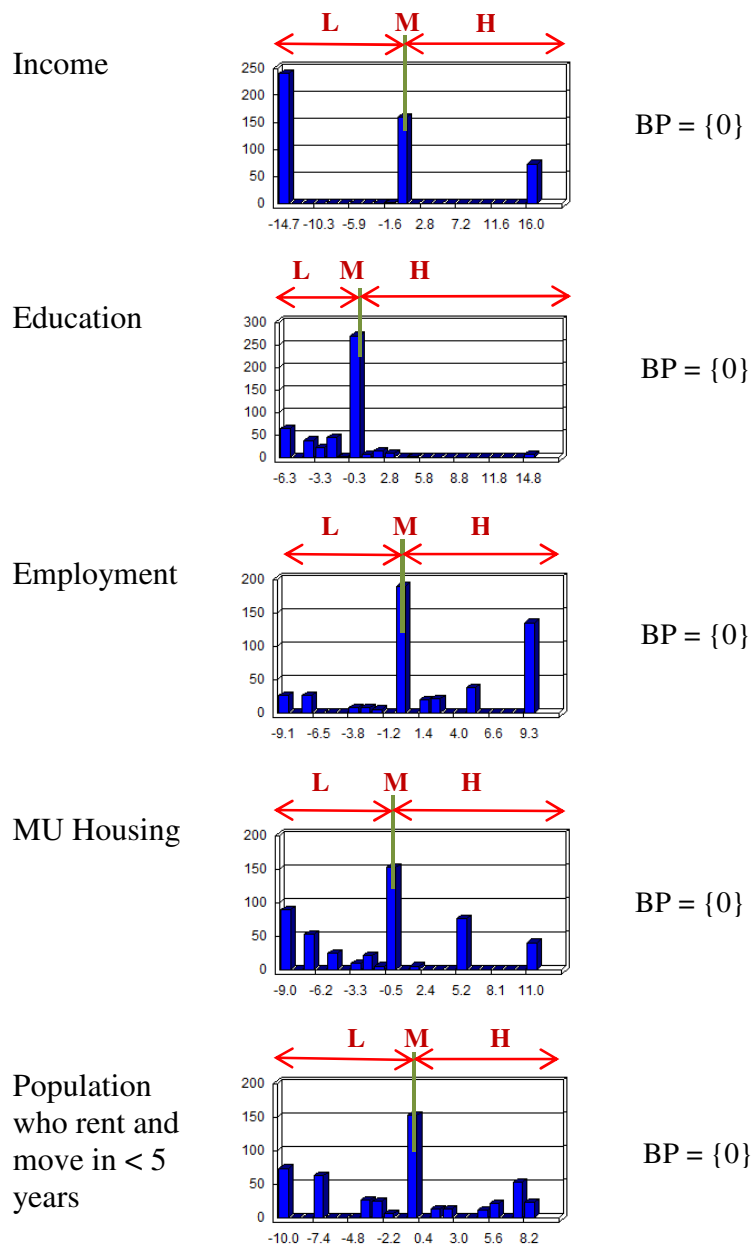
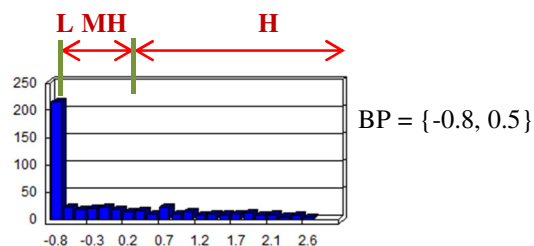


Figure 70: Standardized predicate values histogram and crisp boundary mapping mechanism for Type 1 predicates (Block group variables)

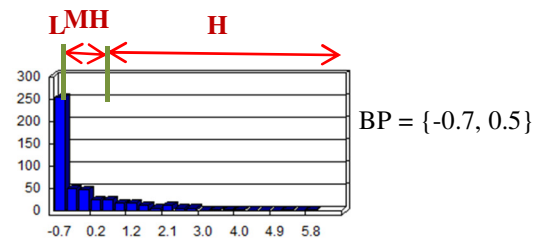
Type 2 predicate – BLK based variables;
 $X = \% \text{ of BLG area has AMOEBA } G^* > 0$

Type 2 predicate – street block based variables;
 $X = \% \text{ of street length has AMOEBA } G^* > 0$

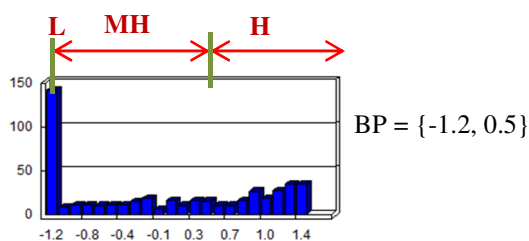
African American



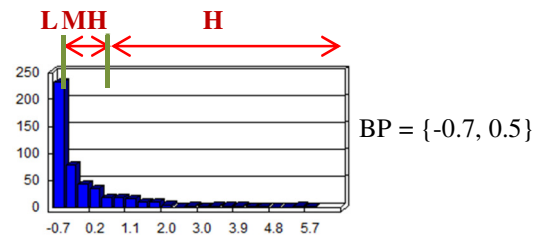
Business



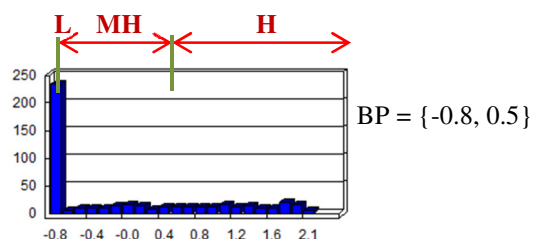
Heterogeneity Index



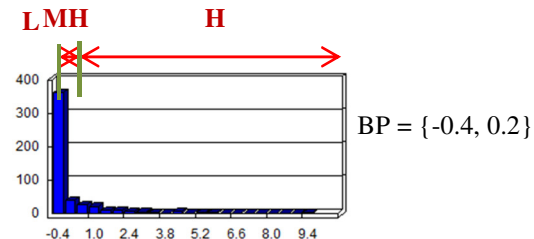
All Crime (CAT)



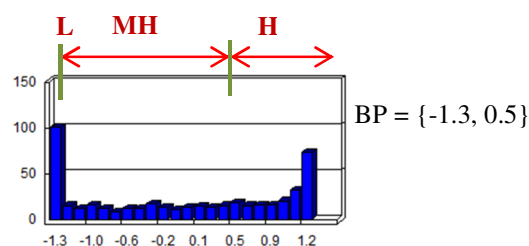
Single parent family



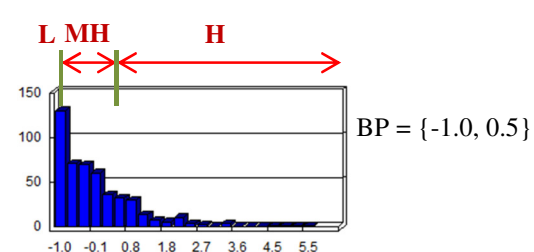
Motor Vehicle Theft (MVT)



Home Owned



Theft from Motor Vehicle (TFM)



Male 17-24

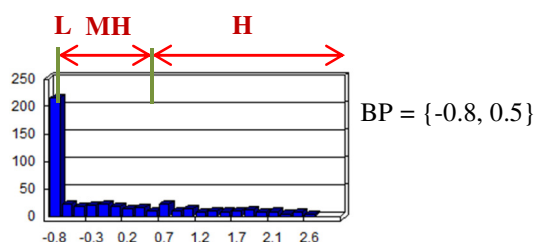


Figure 71: Standardized predicate value histogram and crisp boundary mapping mechanism for Type 2 predicates (Block and Street variables)

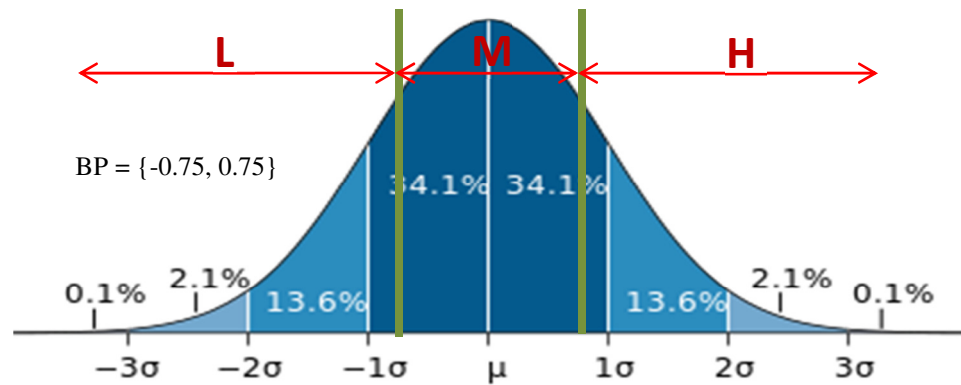


Figure 72: Crisp boundary mapping mechanism for Type 3 predicates (SIM based values) using standardized values

After the predicate values are mapped, the process of generating the final relational table for SAR mining is straightforward. For demonstration purposes, a portion of the final table to mine dangerous street SARs with crisp mapping is shown in Figure 73. The last column of the final table to mine is often dedicated to the decision variable which in this case is either crime of all type, motor vehicle thefts, or thefts from motor vehicle.

BLG ID	Income	HeteInx	Busi	simP Income	simN Income	simP Hete	simP Busi	simPOI DnkPlc	simP CAT	CAT
1	L	MH	H	M	H	L	M	M	L	H
2	M	MH	L	M	H	L	M	L	L	MH
3	M	L	MH	M	H	L	M	M	L	L
4	H	L	L	M	L	H	H	M	H	L
5	L	MH	MH	L	H	L	L	L	L	MH
6	H	L	MH	L	H	L	L	L	L	MH
7	M	L	H	L	H	L	L	L	L	MH
8	L	H	H	M	M	M	M	M	M	H
...

Figure 73: Format of final relational table to mine using crisp mapping

6.4.3 Numeric-to-Nominal Mapping with Fuzzy Boundary

As discussed earlier in Chapter 5, classification using crisp boundary mapping does not have the flexibility to accommodate fuzziness in the linguistic expressions of aspatial and spatial attributes, and thus, fuzzy SAR holds the promise of better discovery performance. In order to provide further examination of this matter, classification with fuzzy boundary mapping will be applied for the case study of mining dangerous streets SAR herein.

A fuzzy mapping mechanism similar to the one presented in Section 5.3.3 is applied for Type 2 and Type 3 predicates as shown in Figure 74 and Figure 75, respectively. Details on membership functions used are reported in Table 11.

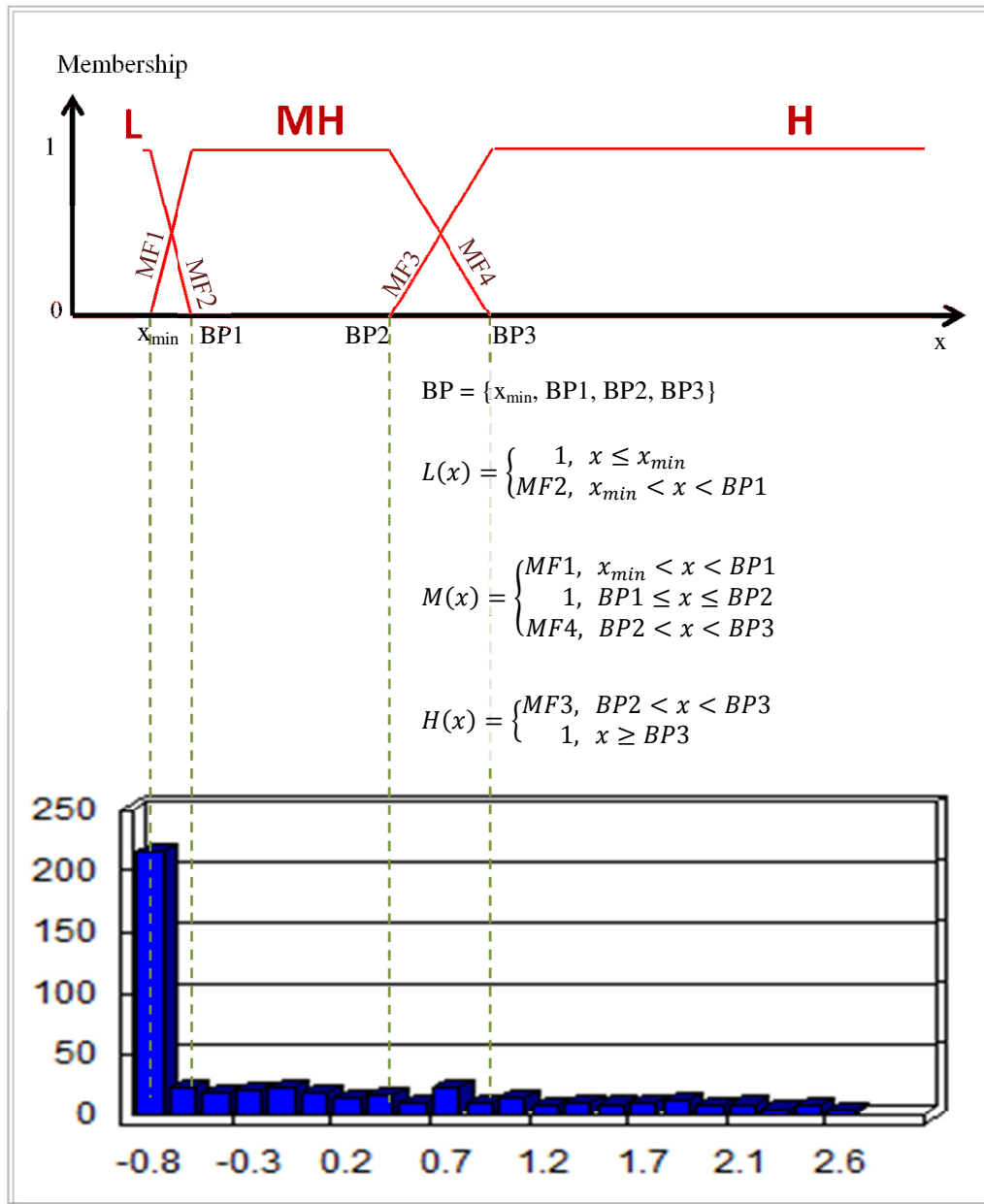


Figure 74: Standardized value histogram and fuzzy boundary mapping mechanism for Type 2 predicates

Table 11: Non-uniform membership functions applied to fuzzy mapping of Type 2 predicates

Predicate	BreakPoint{ }		slope	intercept
African	{-0.80827, -0.5, 0.5, 1}	MF1	-3.24394	-1.62197
American		MF2	3.24394	2.62197
		MF3	-2.00000	2.00000
		MF4	2.00000	-1.00000
Heterogeneity Index	{-1.17184, -0.5, 1, 0}	MF1	-1.48845	-0.74422
		MF2	1.48845	1.74422
		MF3	-2.00000	2.00000
		MF4	2.00000	-1.00000
Male 17-24	{-0.80173, -0.2, 1, 0}	MF1	-1.66188	-0.33238
		MF2	1.66188	1.33238
		MF3	-1.25000	1.25000
		MF4	1.25000	-0.25000
Home Own	{-1.34659, -0.5, 1, 0}	MF1	-1.18121	-0.59061
		MF2	1.18121	1.59061
		MF3	-2.00000	2.00000
		MF4	2.00000	-1.00000
Single-parent family	{-0.83633, -0.5, 1, 0}	MF1	-2.97326	-1.48663
		MF2	2.97326	2.48663
		MF3	-2.00000	2.00000
		MF4	2.00000	-1.00000
CAT	{-0.6549, -0.5, 1, 0}	MF1	-6.45582	-3.22791
		MF2	6.45582	4.22791
		MF3	-2.00000	2.00000
		MF4	2.00000	-1.00000

Busi	{-0.65228, -0.5, 1, 0}	MF1	-6.56689	-3.28345
		MF2	6.56689	4.28345
		MF3	-2.00000	2.00000
		MF4	2.00000	-1.00000
MVT	{-0.41826, -0.2, 1, 0}	MF1	-4.58169	-0.91634
		MF2	4.58169	1.91634
		MF3	-1.25000	1.25000
		MF4	1.25000	-0.25000
TFM	{-0.98886, -0.5, 1, 0}	MF1	-2.04560	-1.02280
		MF2	2.04560	2.02280
		MF3	-2.00000	2.00000
		MF4	2.00000	-1.00000

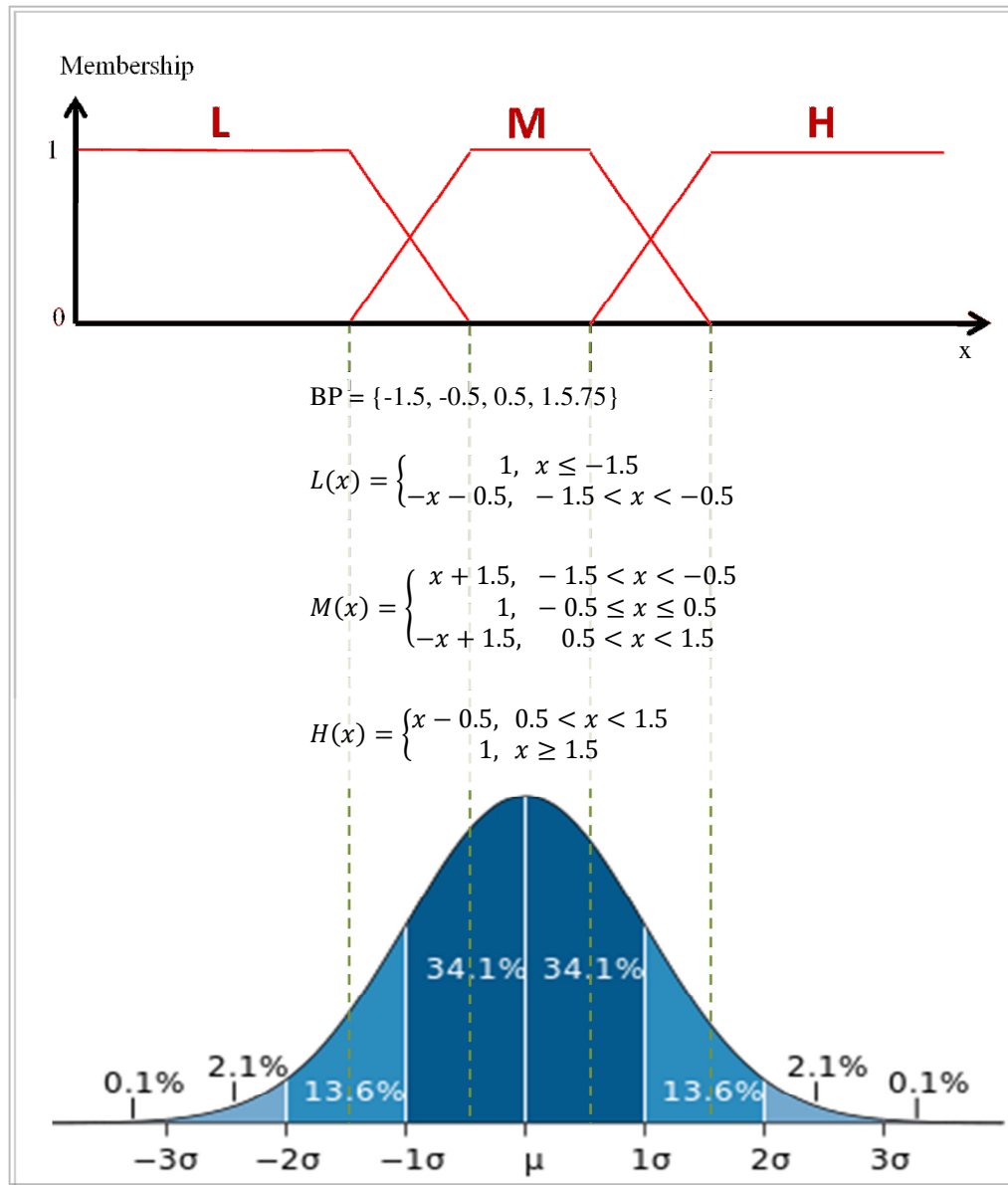


Figure 75: Fuzzy boundary mapping mechanism for Type 3 predicates using standardized values

6.4.4 Ready-to-mine input files

For increasing computing efficiency, all predicates in the final relational table are ID coded with details reported in Table 12. The final relational table is then transformed using the ID codes as shown in Figure 76 and Figure 77, respectively, for crisp and fuzzy SAR mining. In both cases, each record contain only the predicates with membership

function value different than 0 (membership function value for predicates using crisp mapping will only take values of 1 or 0). Using the predicate name and ID schema, the record shown in Figure 76, for instance, will be interpreted as:

{Income=L MUHouse=H HEdu=L UnstableRent=M ... SimPIncome=L
SimNIncome=M ... CAT=H}.

The same record but using fuzzy mapping reported in Figure 77 in addition includes non-zero membership values ranging from [0..1] with each of involving predicates. It is then interpreted as:

{<Income=L, 1> <MUHouse=H, 1> <HEdu=L, 1> <UnstableRent=M, 1> ...
<SimPIncome=M,0.41> <SimPIncome=L,0.59> <SimNIncome=M, 0.8>
<SimNIncome=L, 0.2> ... CAT=H, 0.42> <CAT=M, 0.58>}.

3 4 8 10 14 16 19 22 26 29 33 35 38 42 43 48 50 54 57 60 62 65 69 71 74 78 80 84 87 89
92 95 98 102 104 106

Figure 76: A record in the final ready-to-mine relational table with crisp mapping

<3,1.00> <4,1.00> <8,1.00> <10,1.00> <14,1.00> <16,1.00> <19,1.00> <22,1.00>
<26,1.00> <29,1.00> <32,0.41> <33,0.59> <35,0.8> <36,0.2> <38,0.83> <39,0.17>
<42,1.00> <43,1.00> <47,0.19> <48,0.81> <50,0.99> <51,0.01> <54,1.00>
<56,0.23> <57,0.77> <59,0.46> <60,0.54> <62,1.00> <65,0.85> <66,0.15>
<68,0.02> <69,0.98> <71,1.00> <74,0.69> <75,0.31> <77,0.5> <78,0.5> <80,0.67>
<81,0.33> <83,0.1> <84,0.9> <86,0.4> <87,0.6> <89,0.68> <90,0.32> <92,0.79>
<93,0.21> <95,0.93> <96,0.07> <98,1.00> <101,0.35> <102,0.65> <104,0.55>
<105,0.45> <106,0.42> <107,0.58>.

Figure 77: A record in the final ready-to-mine relational table with fuzzy mapping

Table 12: Predicate name - ID schema table for crisp SAR in the ready-to-mine format

1	Income=H	41	SimN3MUHouse=M	81	SimMall=L
2	Income=M	42	SimN3MUHouse=L	82	SimParkNRi=H
3	Income=L	43	SimPHEdu=H	83	SimParkNRi=M
4	MUHouse=H	44	SimPHEdu=M	84	SimParkNRi=L
5	MUHouse=M	45	SimPHEdu=L	85	SimHSchool=H
6	MUHouse=L	46	SimNHEdu=H	86	SimHSchool=M
7	HEdu=H	47	SimNHEdu=M	87	SimHSchool=L
8	HEdu=M	48	SimNHEdu=L	88	SimCollege=H
9	HEdu=L	49	SimPRMI5L=H	89	SimSimCollege=M
10	UnstableRent=H	50	SimPRMI5L=M	90	SimCollege=L
11	UnstableRent=M	51	SimPRMI5L=L	91	SimDnkPlac=H
12	UnstableRent=L	52	SimNRMI5L=H	92	SimDnkPlac=M
13	Employment=H	53	SimNRMI5L=M	93	SimDnkPlac=L
14	Employment=M	54	SimNRMI5L=L	94	SimWalmart=H
15	Employment=L	55	SimPWork=H	95	SimWalmart=M
16	AfricanA=H	56	SimPWork=M	96	SimWalmart=L
17	AfricanA=MH	57	SimPWork=L	97	Business=H
18	AfricanA=L	58	SimNWork=H	98	Business=MH
19	HeterogeneityInx=H	59	SimNWork=M	99	Business=L
20	HeterogeneityInx=MH	60	SimNWork=L	100	SimCAT=H
21	HeterogeneityInx=L	61	SimHete=H	101	SimCAT=M
22	Male17-24=H	62	SimHete=M	102	SimCAT=L
23	Male17-24=MH	63	SimHete=L	103	SimBusi=H
24	Male17-24=L	64	SimAA=H	104	SimBusi=M
25	HomeOwn=H	65	SimAA=M	105	SimBusi=L
26	HomeOwn=MH	66	SimAA=L	106	CAT=H
27	HomeOwn=L	67	SimHOwn=H	107	CAT=MH
28	SingleParentF=H	68	SimHOwn=M	108	CAT=L
29	SingleParentF=MH	69	SimHOwn=L		
30	SingleParentF=L	70	SimM1724=H		
31	SimPIncome=H	71	SimM1724=M		
32	SimPIncome=M	72	SimM1724=L		
33	SimPIncome=L	73	SimSF=H		
34	SimNIncome=H	74	SimSF=M		
35	SimNIncome=M	75	SimSF=L		
36	SimNIncome=L	76	SimHMTel=H		
37	SimP3MUHouse=H	77	SimHMTel=M		
38	SimP3MUHouse=M	78	SimSimHMTel=L		
39	SimP3MUHouse=L	79	SimMall=H		
40	SimN3MUHouse=H	80	SimMall=M		

6.5 SpatialARMED Level 4: Mining Rules

LUCS-KDD-ARM software package developed by the KDD group at the Department of Computer Science, University of Liverpool is used in this study. Apriori based algorithms for both crisp and fuzzy SAR mining are implemented with fully open-source code which is ideal for any further development or customization. Software downloads and documentation can be found in Coenen (2004).

The process of mining SAR with LUCS-KDD-ARM's Apriori-based algorithm is summarized in Figure 78. This process involves inputting user-defined parameters and data; generating frequent item sets and storing their supports in a tree structure; and generating strong rules with confidence larger than the confidence threshold.

User input on the maximum number of frequent sets (MAX_{FS}) along with support and confidence thresholds (δ_S , δ_C) needs to be application and dataset oriented in order to discover interesting rules. For instance, mining SARs for dangerous streets due to crime in this case study is likely to face the challenge of rare rules, which are interesting rules with very low support and high confidence. When crime incidents are mapped to their corresponding street blocks, only a very small portion (less than 5% in this case study) of the street blocks contains crime records. So one can only mine strong association rules with the consequent indicating *high crime* or *has crime, if any exist*, at support threshold of 5% or less. For a large dataset, the problem surfaces as the algorithm finds too many non-interesting frequent patterns generated at the required low support threshold. Effective pruning strategies are required in order to reveal interesting rare rules. For this particular case study, frequent item sets are generated at a low support threshold (5%). This is to accommodate for the low percentage of streets which have crime. In addition,

the algorithm is modified to only generate frequent rules with respect to high CAT, high MVT, and high TFM by modifying the mining algorithm to only output frequent sets and rules which contain a predicate of high CAT, high MVT, or high TFM. Purposely, this is done to reduce the overwhelming number of output rules while still efficiently serve the purpose of demonstrating and validating SpatialARMED in practice for criminology. It, by no mean, suggests that associations to streets of no or low crime is not interesting. Small modification in the mining algorithm could be made to output these low crime associations. Table 13 entails statistics on frequent sets and rules generated when mining associations to dangerous streets in this study. Due to the large number of rules that have been generated, representative rules are shown in Appendix A to Appendix F for CAT, MVT, and TFM, using crisp and fuzzy mapping.

Table 13: Statistics on generated frequent item sets and rules for dangerous streets SAR mining

	δ_s %	δ_c %	Number of generated frequent item sets	Number of generated rules with high crime as consequent
CAT	5	40	74,476	50,701
MVT	5	40	24,961	10,890
TFM	5	40	75,216	41,333
CAT	5	40	1,508,974	80,075
MVT	5	40	2,124,162	2,503
TFM	5	40	1,501,553	129,908

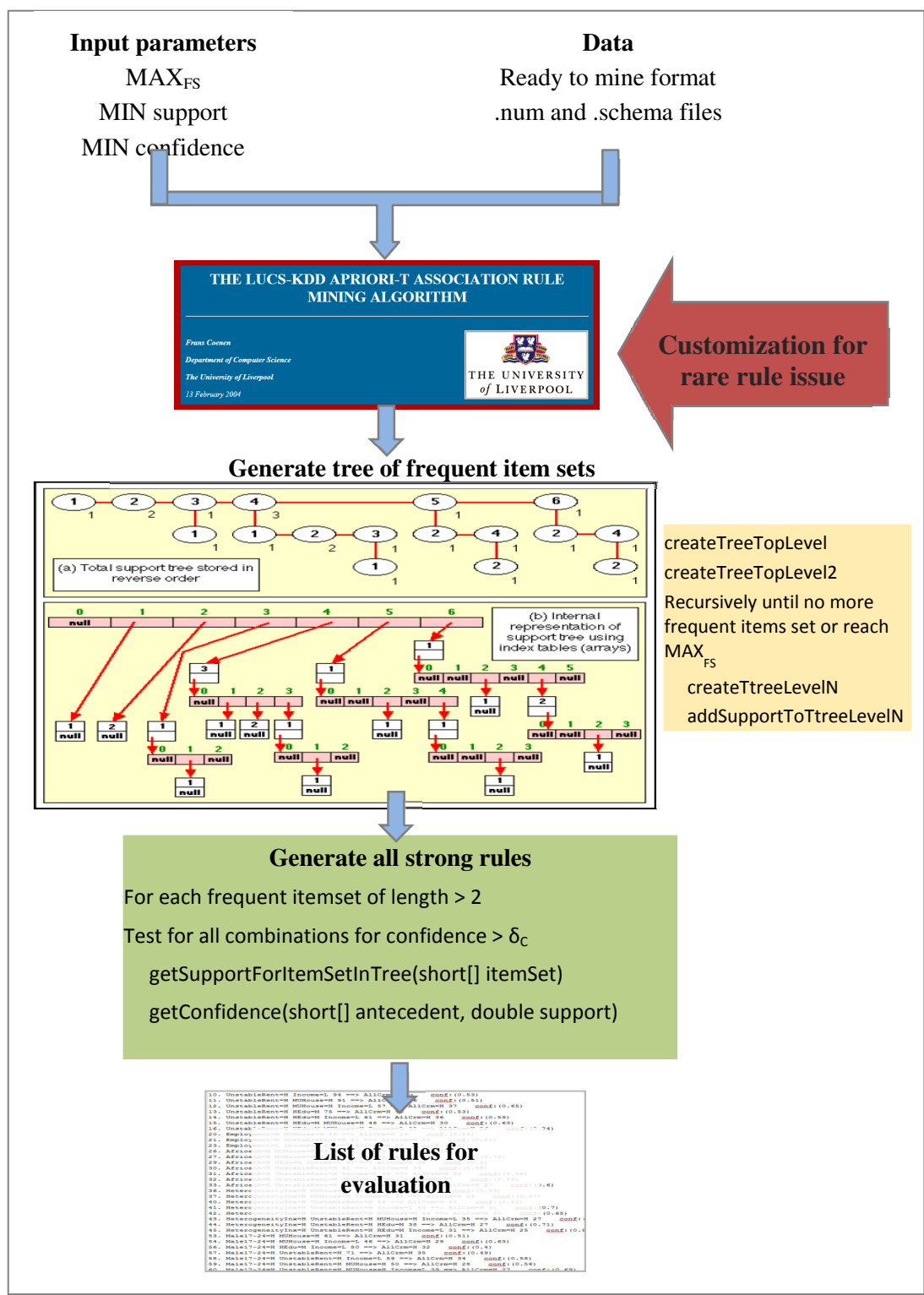


Figure 78: Process of mining SAR with LUCS KDD ARM

6.6 Discovered Associations and SpatialARMED Valuation

The domain knowledge integrated rule evaluation process proposed and discussed in Section 5.4 is applied in this case study to evaluate the above mined results. The very first task of this evaluation process involves the development of a library of known and unknown associations to crime of all types (CAT), motor vehicle thefts (MVT), and thefts from motor vehicles (TFM). Domain experts in criminology can be asked for inputting their knowledge during the construction phase of the library, as discussed in Section 5.4. For this study, the library of known and unknown association to crime is constructed using spatial crime domain knowledge of the author reviewed in Chapter 3. While it is admittedly not the ultimate knowledge, it can serve as a starting point and can be expanded to integrate participations from more domain experts, as discussed in Section 5.4. The library is constructed and shown in Table 14 in its simplest format as a relational table. Known associations here refer to the ones which are well documented by the related body of literature or well acknowledged among domain experts who are evaluating the mining results. Unknown associations are not well documented or related to an uncertainty or controversy in regard as known or unknown within the body of literature. In Table 14, the known associations to crime such as low income, multiple housing structure, low education, unstable rent, high concentration of African-Americans, and high concentration of males aged 17-24, ethnic heterogeneity, and high concentration of businesses are marked with “K” as they are well documented in the literature (reviewed in Section 3.3) as an association to crime. Any other association involved in the mining task is listed as “U” (i.e. unknown). This does not mean that SpatialARMED claim any mined rules, which involve one or more of these so-called unknown

associations, are new or interesting. This K-U definition only serves in representing the integrated existing domain knowledge in the process of rule evaluation, in the effort to sprun the already known (thus non-interesting) associations. One would perhaps want to be more conservative in defining what is known in order to avoid over-prunning. As discussed in Section 5.4, all rules which involved at least one unknown predicate will be classified as *discovering rules* (i.e. rules entering the next loop of evaluation process), which is the focus of futher evaluation based on subgrouping and visual analytics. On the other hand, rules whose predicates are all known will be used as a confirmative measure toward the existing domain knowledge. These rules are therefore referred to as confirmative rules.

Table 14: Library of known associations to crime

Association (Predicate name)	To CAT	To MVT	To TFM
Income_H	U	U	U
Income_M	U	U	U
Income_L	K	K	K
MUHouse_H	K	K	K
MUHouse_M	U	U	U
MUHouse_L	U	U	U
HEdu_H	U	U	U
HEdu_M	U	U	U
HEdu_L	K	K	K
UnstableRent_H	K	K	K
UnstableRent_M	U	U	U
UnstableRent_L	U	U	U
Employment_H	U	U	U
Employment_M	U	U	U
Employment_L	K	K	K
AfricanA_H	K	K	K

AfricanA_MH	K	K	K
AfricanA_L	U	U	U
HeterogeneityInx_H	K	K	K
HeterogeneityInx_MH	K	K	K
HeterogeneityInx_L	U	U	U
Male17-24_H	K	K	K
Male17-24_MH	K	K	K
Male17-24_L	U	U	U
HomeOwn_H	U	U	U
HomeOwn_MH	U	U	U
HomeOwn_L	K	K	K
SingleParentF_H	K	K	K
SingleParentF_MH	K	K	K
SingleParentF_L	U	U	U
SimPIncome_H	U	U	U
SimPIncome_M	U	U	U
SimPIncome_L	U	U	U
SimNIncome_H	U	U	U
SimNIncome_M	U	U	U
SimNIncome_L	U	U	U
SimP3MUHouse_H	U	U	U
SimP3MUHouse_M	U	U	U
SimP3MUHouse_L	U	U	U
SimN3MUHouse_H	U	U	U
SimN3MUHouse_M	U	U	U
SimN3MUHouse_L	U	U	U
SimPHedu_H	U	U	U
SimPHedu_M	U	U	U
SimPHedu_L	U	U	U
SimNHedu_H	U	U	U
SimNHedu_M	U	U	U
SimNHedu_L	U	U	U
SimPRMI5L_H	U	U	U
SimPRMI5L_M	U	U	U
SimPRMI5L_L	U	U	U

SimNRMI5L_H	U	U	U
SimNRMI5L_M	U	U	U
SimNRMI5L_L	U	U	U
SimPWork_H	U	U	U
SimPWork_M	U	U	U
SimPWork_L	U	U	U
SimNWork_H	U	U	U
SimNWork_M	U	U	U
SimNWork_L	U	U	U
SimHete_H	U	U	U
SimHete_M	U	U	U
SimHete_L	U	U	U
SimAA_H	U	U	U
SimAA_M	U	U	U
SimAA_L	U	U	U
SimHOwn_H	U	U	U
SimHOwn_M	U	U	U
SimHOwn_L	U	U	U
SimM1724_H	U	U	U
SimM1724_M	U	U	U
SimM1724_L	U	U	U
SimSF_H	U	U	U
SimSF_M	U	U	U
SimSF_L	U	U	U
SimHMTel_H	U	U	U
SimHMTel_M	U	U	U
SimSimHMTel_L	U	U	U
SimMall_H	U	U	U
SimMall_M	U	U	U
SimMall_L	U	U	U
SimParkNRi_H	U	U	U
SimParkNRi_M	U	U	U
SimParkNRi_L	U	U	U
SimHSchool_H	U	U	U
SimHSchool_M	U	U	U

SimHSchool_L	U	U	U
SimCollege_H	U	U	U
SimSimCollege_M	U	U	U
SimCollege_L	U	U	U
SimDnkPlac_H	U	U	U
SimDnkPlac_M	U	U	U
SimDnkPlac_L	U	U	U
SimWalmart_H	U	U	U
SimWalmart_M	U	U	U
SimWalmart_L	U	U	U
Business_H	K	K	K
Business_MH	U	U	U
Business_L	U	U	U
SimBusi_H	U	U	U
SimBusi_M	U	U	U
SimBusi_L	U	U	U
SimCAT_H	U	U	U
SimCAT_M	U	U	U
SimCAT_L	U	U	U
SimMVT_H	U	U	U
SimMVT_M	U	U	U
SimMVT_L	U	U	U
SimTFM_H	U	U	U
SimTFM_M	U	U	U
SimTFM_L	U	U	U

Following the interactive branching evaluation approach discussed in Section 5.4 with the established library of known and unknown associations, mined rules are classified into two categorized: *confirmative rules* and *discovery rules*. These sets of rules are then evaluated using the domain expert integrated evaluation process proposed in Section 5.4. Interesting findings on the mined SARs for this case study will be discussed for both confirmative rules and discovery rules, one at a time in the following two

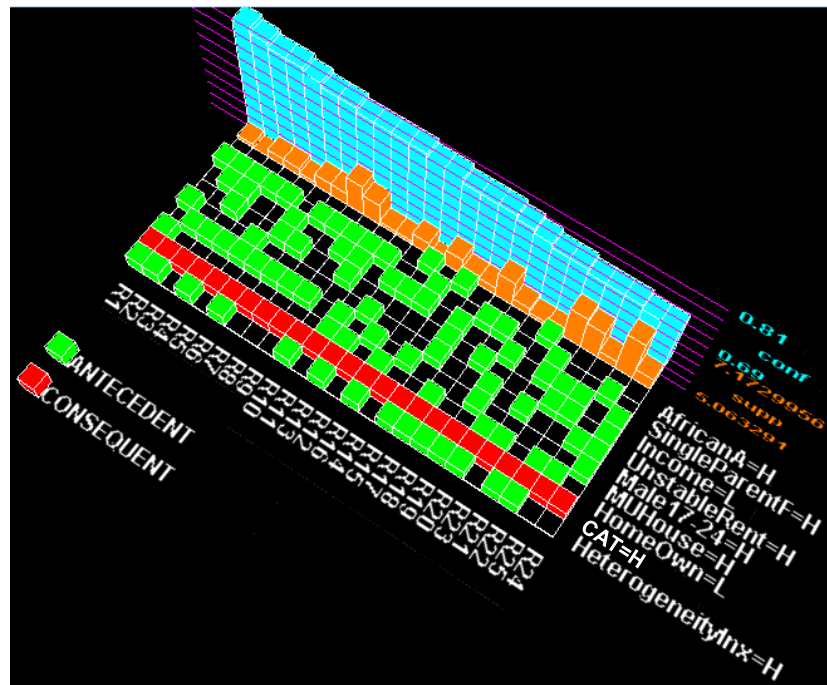
subsections 6.6.1 and 6.6.2. However, it should be emphasized that the purpose of practicing the SpatialARMED framework in criminology for this research is not to report an exhaustive list of interesting rules for crime, but rather to demonstrate the process of SpatialARMED in practice and to evaluate (or validate to some extent) its performance. Guidelines for the performance valuation of SpatialARMED should be in line with its foremost objective proposed in this research, which is to integrate spatial dependence structures embedded in the phenomena under study during the mining process. In order to be recognized for its potentials, SpatialARMED needs to be proven effective, with the case study, in capturing spatial components of crime and the associates, and consequently in discovery rules which not only confirm existing crime patterns but also give further insights leading towards the discovery of potentially new knowledge.

6.6.1 SpatialARMED in Discovering Confirmative Rules

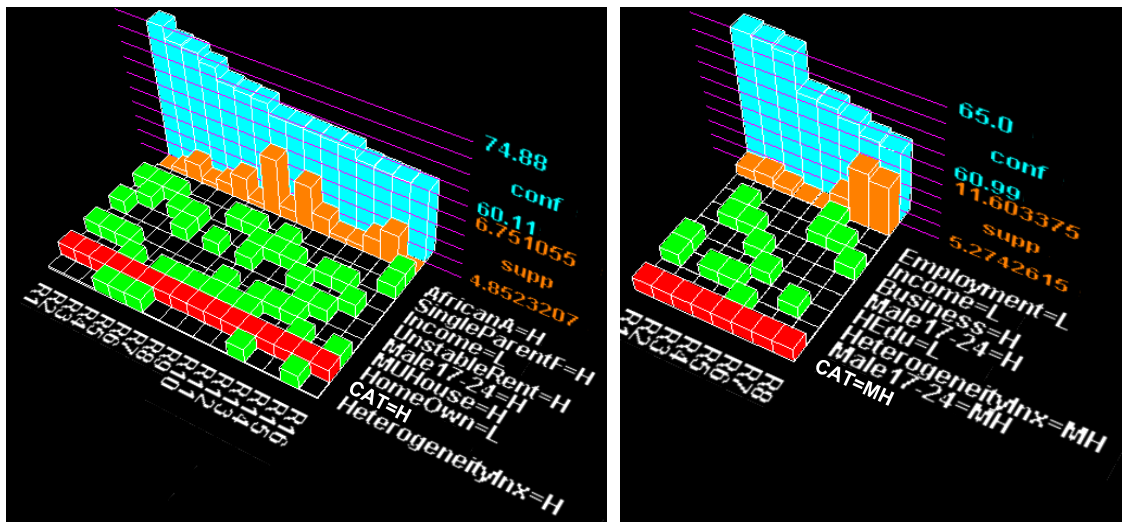
By integrating a data-driven approach (AMOEBA) to capture the spatial dependent structures embedded in the data instead of relying on pre-conceived relationship, the SpatialARMED framework successfully confirms various known frequent associations to crime, either of all types, of motor vehicle thefts, or thefts from motor vehicles for the Charlotte metropolitan area. Besides, it is also successful in evidencing a better performance for SAR mining with fuzzy mapping, instead of with crisp mapping. These accomplishments will be verified through the following discussion on the mined result.

Regarding crime of all types, Table 15 and Table 16, respectively, present the rules of confirmative nature found for high and medium high crime of all types using crisp and fuzzy mapping. Visualizations of these rules are reported in Figure 79. The first twenty-four strongest ones with highest confidence for crisp SAR are shown in Figure 79a and

all mined fuzzy rules are in Figure 79b. SpatialARMED SAR mining using crisp mapping confirms the association of African-American concentration, single-parent family concentration, young male concentration, low income, unstable rent, multiple-unit (MU) housing structure, and ethnic heterogeneity to high crime. SpatialARMED SAR mining on the same dataset but using fuzzy mapping (Figure 79b) advises a better performance by picking up unemployment and high business activity in addition, as associations to medium-high crime. Fuzzy SAR mining also seems to perform better in this case by extracting rules for different levels of crime concentration (i.e. both high and medium high) in comparison to the case of crisp SAR which does not detect any frequent patterns associating with medium-high crime. These indicated associations to crime such as unemployment, low income, unstable rent, multiple unit housing structure, single-parent family concentration, African-American population concentration, high young male population, and ethnic heterogeneity, are consistent with those related to the social disorganization theory and neighborhood effects discussed in Chapter 3. Associations of high business activity and high young male population to crime are very much in line with routine theory and crime pattern theory which relate crime to everyday activities, the physical environment, and the locations of crime generators and attractors. While areas with high business activities often serve as destinations of people's daily activities, they are also often perceived as hubs in term of the criminal cognitive map due to their locations, travel accessibility and containment of crime generators and attractors.



(a)



(b)

Figure 79: Visualization of confirmative SAR for CAT with (a) crisp and (b) fuzzy boundary mapping

Table 15: List of confirmative rules for CAT using crisp boundary mapping

ID	Ancetendent	AntecedentSupport	==>	Consequent	ConsequentSupport	Confidence (decimal %)
1.	SingleParentF=H	HomeOwn=L	Male17-24=H	HeterogeneityInx=H	31	==> CAT=H 25 conf:(0.81)
2.	SingleParentF=H	HeterogeneityInx=H	MUHouse=H	Income=L	30	==> CAT=H 24 conf:(0.8)
3.	SingleParentF=H	UnstableRent=H	MUHouse=H	Income=L	33	==> CAT=H 26 conf:(0.79)
4.	SingleParentF=H	HeterogeneityInx=H	MUHouse=H	33	==> CAT=H 26 conf:(0.79)	
5.	SingleParentF=H	Male17-24=H	MUHouse=H	32	==> CAT=H 25 conf:(0.78)	
6.	HeterogeneityInx=H	UnstableRent=H	MUHouse=H	Income=L	35	==> CAT=H 27 conf:(0.77)
7.	SingleParentF=H	UnstableRent=H	MUHouse=H	35	==> CAT=H 27 conf:(0.77)	
8.	SingleParentF=H	MUHouse=H	41		==> CAT=H 31 conf:(0.76)	
9.	SingleParentF=H	MUHouse=H	Income=L	38	==> CAT=H 29 conf:(0.76)	
10.	SingleParentF=H	HeterogeneityInx=H	UnstableRent=H	Income=L	33	==> CAT=H 25 conf:(0.76)
11.	SingleParentF=H	HomeOwn=L	MUHouse=H	33	==> CAT=H 25 conf:(0.76)	
12.	AfricanA=H	MUHouse=H	Income=L	32	==> CAT=H 24 conf:(0.75)	
13.	HomeOwn=L	Male17-24=H	HeterogeneityInx=H	Income=L	36	==> CAT=H 27 conf:(0.75)
14.	AfricanA=H	MUHouse=H	35		==> CAT=H 26 conf:(0.74)	
15.	Male17-24=H	HeterogeneityInx=H	UnstableRent=H	Income=L	35	==> CAT=H 26 conf:(0.74)
16.	SingleParentF=H	HomeOwn=L	HeterogeneityInx=H	Income=L	38	==> CAT=H 28 conf:(0.74)
17.	SingleParentF=H	HomeOwn=L	HeterogeneityInx=H	41	==> CAT=H 30 conf:(0.73)	
18.	SingleParentF=H	HeterogeneityInx=H	UnstableRent=H	36	==> CAT=H 26 conf:(0.72)	
19.	Male17-24=H	HeterogeneityInx=H	MUHouse=H	Income=L	35	==> CAT=H 25 conf:(0.71)
20.	HomeOwn=L	HeterogeneityInx=H	AfricanA=H	35	==> CAT=H 25 conf:(0.71)	
21.	HeterogeneityInx=H	UnstableRent=H	Income=L	44	==> CAT=H 31 conf:(0.7)	
22.	HomeOwn=L	HeterogeneityInx=H	UnstableRent=H	Income=L	37	==> CAT=H 26 conf:(0.7)
23.	SingleParentF=H	HomeOwn=L	Male17-24=H	Income=L	46	==> CAT=H 32 conf:(0.7)
24.	Male17-24=H	UnstableRent=H	MUHouse=H	Income=L	39	==> CAT=H 27 conf:(0.69)
25.	SingleParentF=H	HomeOwn=L	Male17-24=H	49	==> CAT=H 34 conf:(0.69)	
26.	SingleParentF=H	HomeOwn=L	Male17-24=H	AfricanA=H	38	==> CAT=H 26 conf:(0.68)
27.	HeterogeneityInx=H	MUHouse=H	Income=L	43	==> CAT=H 29 conf:(0.67)	
28.	HomeOwn=L	Male17-24=H	HeterogeneityInx=H	43	==> CAT=H 29 conf:(0.67)	
29.	SingleParentF=H	HomeOwn=L	Male17-24=H	UnstableRent=H	38	==> CAT=H 25 conf:(0.66)
30.	UnstableRent=H	MUHouse=H	Income=L	57	==> CAT=H 37 conf:(0.65)	
31.	HeterogeneityInx=H	UnstableRent=H	MUHouse=H	46	==> CAT=H 30 conf:(0.65)	
32.	HomeOwn=L	HeterogeneityInx=H	Income=L	52	==> CAT=H 34 conf:(0.65)	
33.	HomeOwn=L	Male17-24=H	MUHouse=H	Income=L	37	==> CAT=H 24 conf:(0.65)
34.	HomeOwn=L	Male17-24=H	AfricanA=H	43	==> CAT=H 28 conf:(0.65)	
35.	HomeOwn=L	Male17-24=H	HeterogeneityInx=H	UnstableRent=H	37	==> CAT=H 24 conf:(0.65)
36.	SingleParentF=H	HomeOwn=L	72		==> CAT=H 47 conf:(0.65)	
37.	SingleParentF=H	HomeOwn=L	Income=L	68	==> CAT=H 44 conf:(0.65)	
38.	SingleParentF=H	HomeOwn=L	AfricanA=H	UnstableRent=H	37	==> CAT=H 24 conf:(0.65)

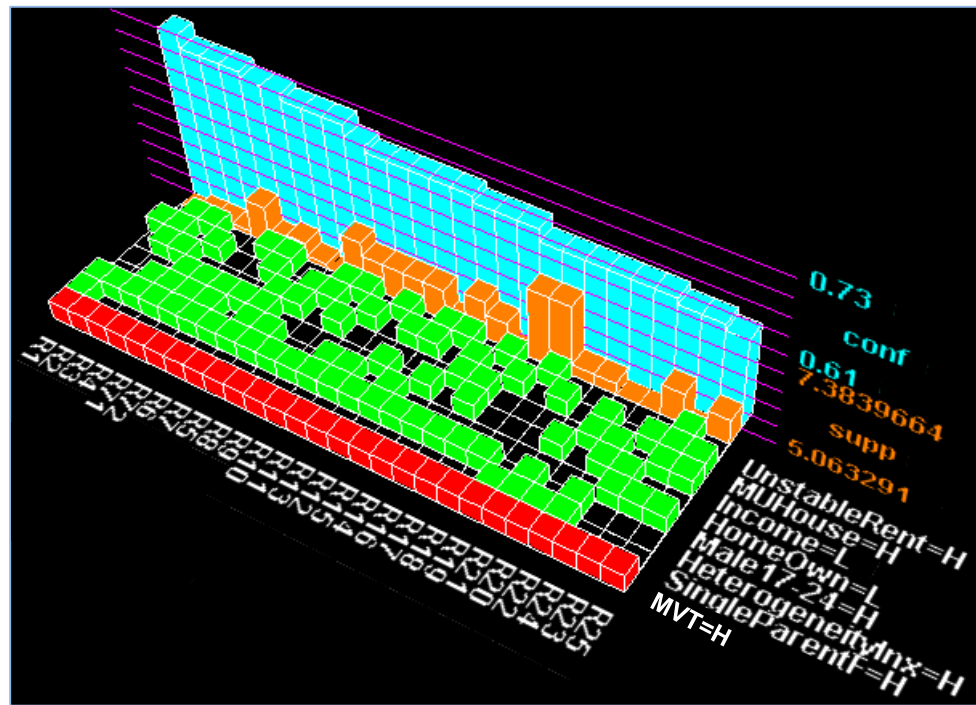
39. HomeOwn=L HeterogeneityInx=H UnstableRent=H MUHouse=H 39 ==> CAT=H 25 conf:(0.64)
 40. HomeOwn=L Male17-24=H AfricanA=H Income=L 42 ==> CAT=H 27 conf:(0.64)
 41. SingleParentF=H HomeOwn=L UnstableRent=H Income=L 47 ==> CAT=H 30 conf:(0.64)
 42. SingleParentF=H HomeOwn=L AfricanA=H 56 ==> CAT=H 36 conf:(0.64)
 43. SingleParentF=H HomeOwn=L AfricanA=H Income=L 55 ==> CAT=H 35 conf:(0.64)
 44. Male17-24=H MUHouse=H Income=L 46 ==> CAT=H 29 conf:(0.63)
 45. Male17-24=H HeterogeneityInx=H UnstableRent=H MUHouse=H 38 ==> CAT=H 24 conf:(0.63)
 46. HomeOwn=L UnstableRent=H MUHouse=H Income=L 48 ==> CAT=H 30 conf:(0.63)
 47. HomeOwn=L HeterogeneityInx=H UnstableRent=H 46 ==> CAT=H 29 conf:(0.63)
 48. SingleParentF=H HomeOwn=L UnstableRent=H 49 ==> CAT=H 31 conf:(0.63)
 49. HomeOwn=L Male17-24=H Income=L 63 ==> CAT=H 39 conf:(0.62)
 50. HeterogeneityInx=H UnstableRent=H 56 ==> CAT=H 34 conf:(0.61)
 51. Male17-24=H HeterogeneityInx=H MUHouse=H 44 ==> CAT=H 27 conf:(0.61)
 52. Male17-24=H HeterogeneityInx=H UnstableRent=H 44 ==> CAT=H 27 conf:(0.61)
 53. HomeOwn=L HeterogeneityInx=H MUHouse=H 44 ==> CAT=H 27 conf:(0.61)
 54. SingleParentF=H Male17-24=H UnstableRent=H Income=L 44 ==> CAT=H 27 conf:(0.61)
 55. HomeOwn=L AfricanA=H 68 ==> CAT=H 41 conf:(0.6)
 56. HomeOwn=L AfricanA=H UnstableRent=H 45 ==> CAT=H 27 conf:(0.6)
 57. SingleParentF=H UnstableRent=H Income=L 62 ==> CAT=H 37 conf:(0.6)
 58. SingleParentF=H Male17-24=H UnstableRent=H 47 ==> CAT=H 28 conf:(0.6)
-

Table 16: Confirmative rules for CAT using fuzzy boundary mapping

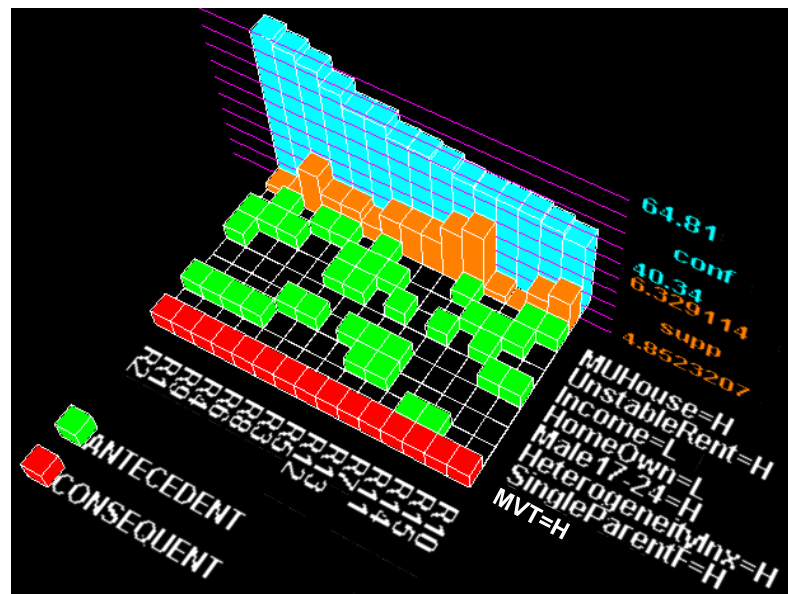
No.	Ancetudent	AntecedentSupport	==>	Consequent	ConsequentSupport	Confidence (%)
1.	SingleParentF_H	UnstableRent_H MUHouse_H 31.89	==>	CAT_H 23.88	conf:(74.88)	
2.	SingleParentF_H	MUHouse_H Income_L 34.55	==>	CAT_H 24.91	conf:(72.08)	
3.	SingleParentF_H	MUHouse_H 37.55	==>	CAT_H 26.91	conf:(71.65)	
4.	HomeOwn_L	HeterogeneityInx_H Income_L 34.55	==>	CAT_H 24.1	conf:(69.74)	
5.	HeterogeneityInx_H	UnstableRent_H Income_L 38.51	==>	CAT_H 25.73	conf:(66.82)	
6.	HomeOwn_L	HeterogeneityInx_H 41.27	==>	CAT_H 27.34	conf:(66.25)	
7.	SingleParentF_H	HomeOwn_L UnstableRent_H 37.07	==>	CAT_H 24.44	conf:(65.93)	
8.	HeterogeneityInx_MH	HEdu_L 47.92	==>	CAT_MH 31.15	conf:(65.0)	
9.	Business_H	Income_L 50.57	==>	CAT_MH 32.8	conf:(64.85)	
10.	Business_H	HeterogeneityInx_MH 44.87	==>	CAT_MH 29.02	conf:(64.67)	
11.	SingleParentF_H	HomeOwn_L 50.78	==>	CAT_H 32.19	conf:(63.38)	
12.	HomeOwn_L	UnstableRent_H MUHouse_H Income_L 39.48	==>	CAT_H 25.01	conf:(63.33)	
13.	SingleParentF_H	HomeOwn_L Income_L 47.77	==>	CAT_H 30.17	conf:(63.16)	
14.	HomeOwn_L	MUHouse_H Income_L 42.51	==>	CAT_H 26.8	conf:(63.03)	
15.	HeterogeneityInx_H	UnstableRent_H MUHouse_H 39.36	==>	CAT_H 24.74	conf:(62.86)	
16.	Male17-24_H	HeterogeneityInx_MH 42.27	==>	CAT_MH 26.38	conf:(62.41)	
17.	Male17-24_MH	HEdu_L 41.3	==>	CAT_MH 25.74	conf:(62.32)	
18.	Employment_L	Income_L 60.0	==>	CAT_MH 37.1	conf:(61.83)	
19.	HeterogeneityInx_MH	Income_L 90.26	==>	CAT_MH 55.24	conf:(61.2)	
20.	Business_H	86.14	==>	CAT_MH 52.54	conf:(60.99)	
21.	Male17-24_H	MUHouse_H Income_L 40.03	==>	CAT_H 24.37	conf:(60.86)	
22.	HomeOwn_L	Male17-24_H Income_L 44.46	==>	CAT_H 26.93	conf:(60.56)	
23.	HeterogeneityInx_H	UnstableRent_H 47.43	==>	CAT_H 28.63	conf:(60.37)	
24.	SingleParentF_H	HomeOwn_L AfricanA_H 39.53	==>	CAT_H 23.76	conf:(60.11)	

Regarding motor vehicle thefts (MVT), Table 17 and Table 18, respectively, present the rules of confirmative nature found for high and medium-high MVT using crisp and fuzzy mapping. Visualizations of the strongest rules with highest confidence for both cases are presented in Figure 80. SAR mining using both crisp and fuzzy boundary mapping confirm associations of unstable rent, multiple-unit housing structure, low

income, young male concentration, single parent family concentration, and heterogeneity to high motor vehicle thefts.



(a)



(b)

Figure 80: Visualization of confirmative SAR for MVT with (a) crisp and (b) fuzzy boundary mapping

Table 17: Listed of confirmative rules for MVT using crisp boundary mapping

No.	Ancetudent	AntecedentSupport	==>	Consequent	ConsequentSupport	Confidence (decimal %)
1.	SingleParentF=H	HeterogeneityInx=H MUHouse=H 33	==>	MVT=H 24	conf:(0.73)	
2.	HeterogeneityInx=H	UnstableRent=H MUHouse=H Income=L 35	==>	MVT=H 25	conf:(0.71)	
3.	Male17-24=H	HeterogeneityInx=H UnstableRent=H Income=L 35	==>	MVT=H 25	conf:(0.71)	
4.	HomeOwn=L	Male17-24=H HeterogeneityInx=H MUHouse=H 35	==>	MVT=H 25	conf:(0.71)	
5.	Male17-24=H	HeterogeneityInx=H MUHouse=H Income=L 35	==>	MVT=H 24	conf:(0.69)	
6.	HomeOwn=L	HeterogeneityInx=H UnstableRent=H MUHouse=H 39	==>	MVT=H 27	conf:(0.69)	
7.	HomeOwn=L	Male17-24=H HeterogeneityInx=H Income=L 36	==>	MVT=H 25	conf:(0.69)	
8.	HeterogeneityInx=H	UnstableRent=H MUHouse=H 46	==>	MVT=H 31	conf:(0.67)	
9.	HeterogeneityInx=H	UnstableRent=H Income=L 44	==>	MVT=H 29	conf:(0.66)	
10.	Male17-24=H	HeterogeneityInx=H MUHouse=H 44	==>	MVT=H 29	conf:(0.66)	
11.	Male17-24=H	HeterogeneityInx=H UnstableRent=H 44	==>	MVT=H 29	conf:(0.66)	
12.	Male17-24=H	HeterogeneityInx=H UnstableRent=H MUHouse=H 38	==>	MVT=H 25	conf:(0.66)	
13.	HomeOwn=L	HeterogeneityInx=H MUHouse=H 44	==>	MVT=H 29	conf:(0.66)	
14.	HeterogeneityInx=H	MUHouse=H Income=L 43	==>	MVT=H 28	conf:(0.65)	
15.	HomeOwn=L	HeterogeneityInx=H UnstableRent=H 46	==>	MVT=H 30	conf:(0.65)	
16.	HomeOwn=L	HeterogeneityInx=H UnstableRent=H Income=L 37	==>	MVT=H 24	conf:(0.65)	
17.	HeterogeneityInx=H	MUHouse=H 56	==>	MVT=H 35	conf:(0.63)	
18.	HeterogeneityInx=H	UnstableRent=H 56	==>	MVT=H 35	conf:(0.63)	
19.	SingleParentF=H	MUHouse=H 41	==>	MVT=H 26	conf:(0.63)	
20.	SingleParentF=H	MUHouse=H Income=L 38	==>	MVT=H 24	conf:(0.63)	
21.	SingleParentF=H	HomeOwn=L HeterogeneityInx=H 41	==>	MVT=H 26	conf:(0.63)	
22.	SingleParentF=H	HomeOwn=L HeterogeneityInx=H Income=L 38	==>	MVT=H 24	conf:(0.63)	
23.	Male17-24=H	UnstableRent=H MUHouse=H Income=L 39	==>	MVT=H 24	conf:(0.62)	
24.	HomeOwn=L	Male17-24=H MUHouse=H 47	==>	MVT=H 29	conf:(0.62)	
25.	Male17-24=H	MUHouse=H Income=L 46	==>	MVT=H 28	conf:(0.61)	

Table 18: Listed of confirmative rules for MVT using fuzzy boundary mapping

No.	Ancetendent	AntecedentSupport	==>	Consequent	ConsequentSupport	Confidence (%)
1.	HeterogeneityInx_H UnstableRent_H MUHouse_H	39.36	==>	MVT_H	24.74	conf:(62.84)
2.	HeterogeneityInx_H UnstableRent_H Income_L	38.51	==>	MVT_H	24.96	conf:(64.81)
3.	UnstableRent_H MUHouse_H Income_L	57.0	==>	MVT_H	27.24	conf:(47.78)
4.	HeterogeneityInx_H MUHouse_H	48.4	==>	MVT_H	26.95	conf:(55.68)
5.	Male17-24_H UnstableRent_H	58.71	==>	MVT_H	27.87	conf:(47.46)
6.	Male17-24_H MUHouse_H	50.75	==>	MVT_H	26.11	conf:(51.45)
7.	MUHouse_H Income_L	70.0	==>	MVT_H	30.41	conf:(43.44)
8.	Male17-24_H UnstableRent_H Income_L	49.48	==>	MVT_H	25.45	conf:(51.42)
9.	HeterogeneityInx_H UnstableRent_H	47.43	==>	MVT_H	28.63	conf:(60.35)
10.	HomeOwn_L MUHouse_H	67.87	==>	MVT_H	27.39	conf:(40.34)
11.	SingleParentF_H UnstableRent_H	57.49	==>	MVT_H	24.91	conf:(43.32)
12.	Male17-24_H HeterogeneityInx_H Income_L	60.09	==>	MVT_H	27.74	conf:(46.16)
13.	Male17-24_H HeterogeneityInx_H	67.68	==>	MVT_H	29.88	conf:(44.14)
14.	SingleParentF_H UnstableRent_H Income_L	55.44	==>	MVT_H	23.91	conf:(43.12)
15.	HomeOwn_L UnstableRent_H MUHouse_H	62.26	==>	MVT_H	26.15	conf:(42.0)

Regarding thefts from motor vehicles (TFM), Table 19 and Table 20, respectively, present the rules of confirmative nature found for high and medium-high TFM using crisp and fuzzy mapping. All mined crisp rules are visualized in Figure 81a and the twenty-four strongest ones with highest confidence for fuzzy SAR are in Figure 81b. SpatialARMED SAR mining using crisp boundary mapping once again confirms associations of unstable rent, multiple-unit housing concentration, young male concentration, low home ownership, single-parent family concentration, and heterogeneity to thefts from motor vehicles. SpatialARMED fuzzy SAR mining in addition confirms associations of African-American concentration, low income, high business, and low employment to theft from motor vehicles.

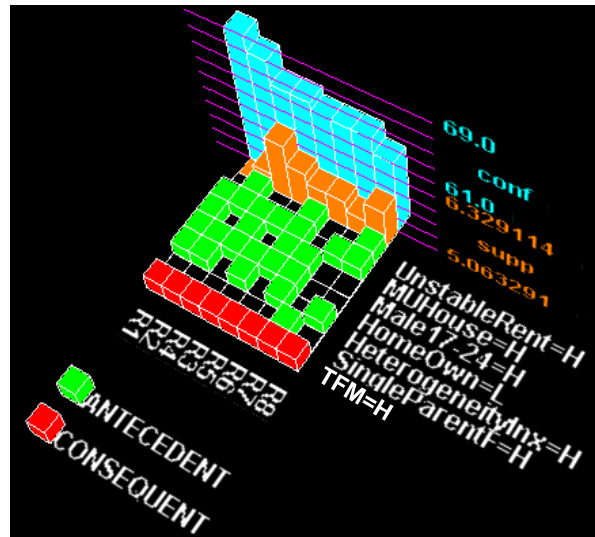
Table 19: Listed of all confirmative rules for thefts from motor vehicle using crisp boundary mapping

No.	Ancetudent	AntecedentSupport	==>	Consequent	ConsequentSupport	Confidence (%)
1.	HomeOwn=L Male17-24=H HeterogeneityInx=H MUHouse=H	35	==>	TFM=H	24	conf:(69)
2.	HomeOwn=L HeterogeneityInx=H UnstableRent=H MUHouse=H	39	==>	TFM=H	26	conf:(67)
3.	HomeOwn=L HeterogeneityInx=H MUHouse=H	44	==>	TFM=H	28	conf:(64)
4.	HomeOwn=L Male17-24=H MUHouse=H	47	==>	TFM=H	30	conf:(64)
5.	HomeOwn=L Male17-24=H UnstableRent=H MUHouse=H	43	==>	TFM=H	27	conf:(63)
6.	HomeOwn=L Male17-24=H HeterogeneityInx=H	43	==>	TFM=H	27	conf:(63)
7.	SingleParentF=H MUHouse=H	41	==>	TFM=H	26	conf:(63)
8.	HeterogeneityInx=H UnstableRent=H MUHouse=H	46	==>	TFM=H	28	conf:(61)

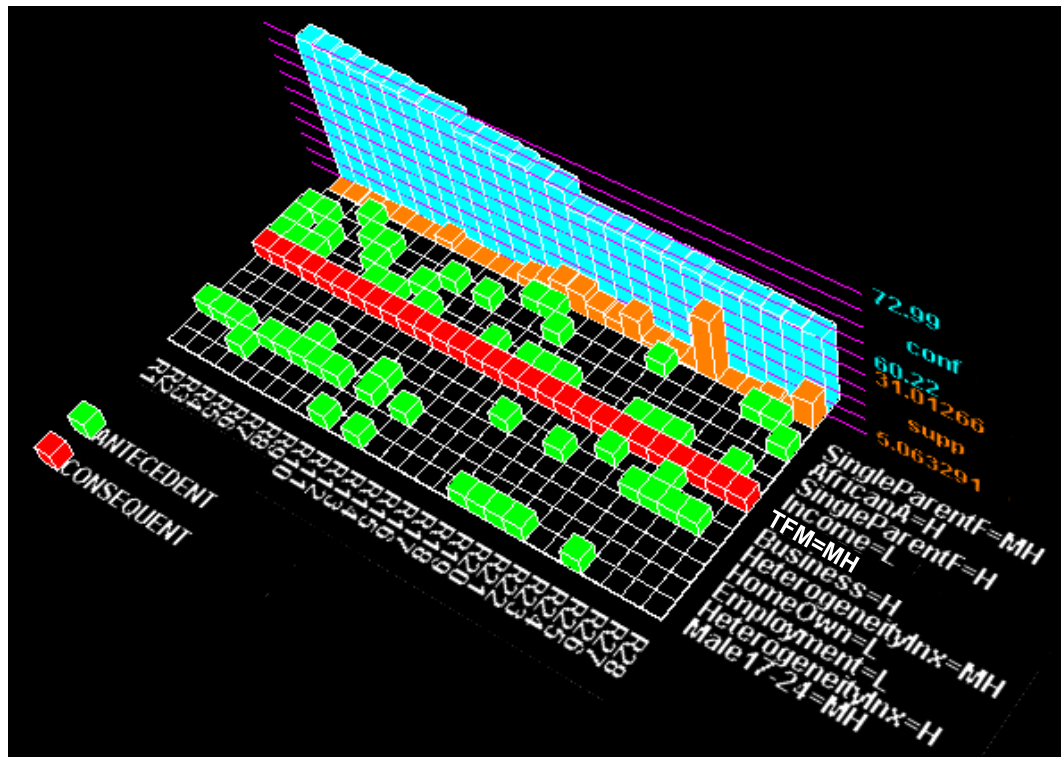
Table 20: Listed of all confirmative rules for thefts from motor vehicle using fuzzy boundary mapping

No.	Ancetudent	AntecedentSupport	==>	Consequent	ConsequentSupport	Confidence (%)
1.	SingleParentF_H AfricanA_H Employment_L Income_L	33.59	==>	TFM_MH	24.52	conf:(72.99)
2.	AfricanA_H Employment_L Income_L	40.64	==>	TFM_MH	29.52	conf:(72.64)
3.	SingleParentF_H Employment_L Income_L	37.63	==>	TFM_MH	27.13	conf:(72.1)
4.	SingleParentF_MH HeterogeneityInx_H	36.91	==>	TFM_MH	26.37	conf:(71.44)
5.	AfricanA_H Employment_L	41.86	==>	TFM_MH	29.83	conf:(71.26)
6.	SingleParentF_H AfricanA_H Employment_L	34.5	==>	TFM_MH	24.52	conf:(71.06)
7.	HomeOwn_L Employment_L Income_L	35.31	==>	TFM_MH	24.89	conf:(70.49)
8.	Employment_L Income_L	60.0	==>	TFM_MH	42.0	conf:(70.0)
9.	SingleParentF_H Employment_L	38.76	==>	TFM_MH	27.13	conf:(70.0)
10.	Male17-24_MH AfricanA_H Income_L	38.11	==>	TFM_MH	26.03	conf:(68.3)
11.	HomeOwn_L Employment_L	36.65	==>	TFM_MH	24.89	conf:(67.91)
12.	Male17-24_MH AfricanA_H	41.87	==>	TFM_MH	28.38	conf:(67.79)
13.	Employment_L	72.0	==>	TFM_MH	48.45	conf:(67.29)
14.	SingleParentF_MH Income_L	84.13	==>	TFM_MH	56.02	conf:(66.58)
15.	SingleParentF_MH	114.89	==>	TFM_MH	75.92	conf:(66.08)
16.	SingleParentF_H MUHouse_H	37.55	==>	TFM_H	24.0	conf:(63.91)
17.	AfricanA_H Income_L	117.53	==>	TFM_MH	73.74	conf:(62.74)

18. HeterogeneityInx_MH Income_L 90.26 ==> TFM_MH 56.53 conf:(62.63)
 19. Male17-24_MH Income_L 81.38 ==> TFM_MH 50.91 conf:(62.56)
 20. Male17-24_MH 120.72 ==> TFM_MH 75.26 conf:(62.34)
 21. Male17-24_MH HeterogeneityInx_MH 60.14 ==> TFM_MH 37.46 conf:(62.28)
 22. SingleParentF_MH Male17-24_MH 49.03 ==> TFM_MH 30.45 conf:(62.1)
 23. Business_H Income_L 50.57 ==> TFM_MH 31.25 conf:(61.79)
 24. Income_L 242.0 ==> TFM_MH 147.89 conf:(61.11)
 25. Male17-24_MH HeterogeneityInx_MH Income_L 42.82 ==> TFM_MH 26.17
conf:(61.11)
 26. Business_H HeterogeneityInx_MH 44.87 ==> TFM_MH 27.4 conf:(61.07)
 27. SingleParentF_MH HeterogeneityInx_MH 60.14 ==> TFM_MH 36.61 conf:(60.88)
 28. SingleParentF_MH HeterogeneityInx_MH Income_L 44.03 ==> TFM_MH 26.65
conf:(60.52)
 29. AfricanA_H 132.93 ==> TFM_MH 80.06 conf:(60.22)
-



(a)



(b)

Figure 81: Visualization of all confirmative SAR for TFM with (a) crisp and (b) fuzzy boundary mapping

For summary and comparison purposes, Table 21 lists all of the known associations to crime that SpatialARMED picked up. It is clear that many of the associations, 13 out

of 15, are confirmed herein by SpatialARMED for CAT, MVT, or TFM. The two associations which are not picked up by SpatialARMED for this particular data set are “AfricanA_MH” and “SingleParentF_MH”. This simply means that, with this dataset, block groups having only a portion of their area having high concentration of African American or of single parent family are not found to be associated with high CAT, MVT, or TFM. Interestingly, out of all mined rules, SpatialARMED does not suggest a strong association of high business, of unemployment, or of African-American concentration to motor vehicle thefts.

Table 21: Confirmed associations to crime by SpatialARMED

Crime of all types	Motor vehicle theft	Theft from motor vehicle
SingleParentF_H	SingleParentF_H	SingleParentF_H SingleParentF_MH
UnstableRent_H	UnstableRent_H	
MUHouse_H	MUHouse_H	MUHouse_H
Income_L	Income_L	Income_L
HomeOwn_L	HomeOwn_L	HomeOwn_L
HeterogeneityInx_H HeterogeneityInx_MH	HeterogeneityInx_H	HeterogeneityInx_H HeterogeneityInx_MH
Male17-24_H Male17-24_MH	Male17-24_H	Male17-24_H
Business_H		Business_H
Employment_L		Employment_L
AfricanA_H		AfricanA_H
HEdu_L		

6.6.2 SpatialARMED in Discovering Potentially New Rules

In addition to confirming existing knowledge, the capability to discovery potentially new ones is crucial to the success of a new data mining framework such as SpatialARMED. In order to evaluate this capability, the rule evaluation processes in

support for new knowledge discovery with both top-down and interactive branching approaches proposed in Section 5.4. are applied on the set of SpatialARMED mined rules for crime. With the top-down approach, the process starts from the strongest rules; on the other hand, with the interactive branching approach, it does from the set of discovery rules, i.e. rules containing at least one unknown predicate. By the first approach, rules with highest confidence and satisfying the support threshold are examined for interesting associations. This approach in practice could face limitation if the number of strong rules is very large and the interesting associations are relatively rare, which means that they cannot be revealed among the first few strong rules. By the second approach, an automatic extraction of the non-repeated predicates involved in the discovery rules, referred here as *discovering predicates*, is performed. The analyst can then use domain knowledge to surf through these predicates and make selections, iteratively, to subgroup the rules in order to facilitate the exploration for discovering new knowledge.

Certainly, some could question the legitimacy of any so-called “potentially new” association indicated by SpatialARMED mining as discussed below. Please do keep in mind that discovery is an iterative process of applying and improving mining algorithms along with the knowledge base integration. In a geographical knowledge discovery process, SpatialARMED mining is thus only one step, whose returned results are tied to one stage of the mining logic (i.e. the algorithm’s brain), in this case, defined by the specifications on the involved semantics and spatial components as well as the library of known and unknown associations. In the next stage(s), these specifications, particularly the library of known and unknown associations, could be iteratively adapted in support for the discovery. Having this as a general context, evaluation on SpatialARMED mining

results presented in this section serve (1) to demonstrate the strength of SpatialARMED mining algorithm by integrating spatial components and, (2) to build up a framework with which mining analysts and domain experts can integratedly work together toward an effective geographical knowledge discovery process.

Meticulously, the novelty of SpatialARMED is at detection and integration of spatial components during the SAR mining process by performing AMOEBA spatial clustering and spatial spillover impact modeling for the detected clusters. This novelty allows SpatialARMED to outperform the traditional spatial analysis techniques such as regression as well as existing AR and SAR mining approaches in discovering associations as the result of both direct and indirect spatial functional relations. Particularly for crime analysis, the direct spatial functional relations herein refer to what have been known from literature as neighborhood effects. SpatialARMED achieves success from this perspective by capturing the spatial dependency structure (i.e. hot and cold spots) of variables under examination and mine associations of crime with respect to these clusters per se. As the result, various well acknowledged neighborhood effects in relation to CAT, MVT, and TFM have been extracted as discussed in previous subsection.

Examination and evaluation on the set of discovery rules as discussed below not only further evidence this success but also suggest SpatialARMED's potential capability in mining indirect spatial functional relations as well as interactions among associating factors, which could be both spatially and nonspatially, to crime. The indirect spatial functional relations are those picked up as the result of capturing the spillover effects of the identified spatial dependence structures (i.e. clusters) on the surroundings. So this is

related to what have been known in the spatial analysis or spatial econometric as proximity or spillover effects. An example of such association is “if a block group has high average income and it is *under strong influence of* multiple unit rental communities, then it likely experiences high thefts from motor vehicle”. The concept of *under strong influence* is defined by the spillover effect models as previously discussed, which could mean "close-to" in the simplest cases, or could involve more complex spatial interaction components. The interactions among associating factors refer to cases where two of more associations are picked up by SpatialARMED in combinations as frequent and strong pattern to crime. Given that these associating factors in a more general context could either present semantic or direct/indirect spatial relations to the phenomenon under study, their combinations indeed represent possible interactive, i.e. compounding or mitigating, effects on the phenomenon. It would be fair to partially credit AR mining inherently for this potential. The nature of AR mining indeed aims to extract all possible combinations of mining predicates so that the criteria of being frequent and strong are satisfied. As the result, suggested rules at varying level of complexity, i.e. number of involving predicates, are returned. Among these, the complex rules, i.e. long rules with many predicates, are particularly valuable as they represent the compounding or mitigating (i.e. interactive) effects created by the involving factors which could be semantically and/or spatially related to the phenomenon under study. Although complex rules often possess lower support in comparison with more simple rules, they may be hold at a high confidence level and remain to be especially interesting. The branching evaluation approach proposed for SpatialARMED in particular allows rules to be interactively classified and

analysed in an effective manner so that complex and strong rules are revealed to provide valuable insights.

The following subsections discuss the results of evaluating SpatialARMED discovery rules from crime, following both top-down and interactive branching evaluation approaches.

6.6.2.1 Examination based on Top-down Approach

Following a top-down evaluation approach, rules are sorted based on their support and confidence values. Rule evaluation are then based on the ones which are most frequent (i.e. having highest support) and strongest (i.e. having highest confidence). Generally, the results continue to confirm various known associations to high crime and to motor vehicle related crime. In addition, complex associations due to neighbourhood effects and/or proximity effects are also successfully extracted.

For high and medium-high CAT, the strongest mined rules are listed in Table 22 and visualized in Figure 82. SpatialARMED suggests that, at 80% confidence, if an area is a rental or multiple-unit housing community and has high percentage of single-parent family then it is associated with highest crime. Interestingly, some sorts of combined spillover effects at medium level of business activity, of unemployment and alcoholic drinking (Rule 1 and Rule 5), or of high schools and low education (Rule 2), or of employment and young male, in addition to high multiple-unit housing and high single parent family, are found associated to high crime. In contrast, residential areas with medium-high homeownership are also found at very high level of confidence (near 80% confidence) to have an association with medium-high crime. However, crime in these medium-high homeownerhip areas is often associated with high business activity (Rule 7

and Rule 8). If high business is not present, then either low income with spillover effects at medium level of rental multiple-unit housing communities (Rule 9) or medium employment with spillover effects at medium level of unstable rent and alcoholic drinking places (Rule 10) are associated to crime in these residential areas. Rule 10 in Figure 82b also indicates if a residential area of medium high homeownership but has high African-American population with spillover impact at medium level of unemployment then it also experiences medium high crime.

Table 22: SpatialARMED strongest rules for CAT

For high CAT

1. SimBusi_M SimNWork_M SingleParentF_H HomeOwn_L 29.87 ==> CAT_H 23.78
conf:(79.62)
2. SimHSchool_M SimNHEdu_M SingleParentF_H MUHouse_H 31.15 ==> CAT_H 24.47
conf:(78.55)
3. SimCollege_M SimNWork_M SingleParentF_H HomeOwn_L 30.3 ==> CAT_H 23.8
conf:(78.54)
4. SimM1724_M SimPWork_M SingleParentF_H HomeOwn_L 30.47 ==> CAT_H 23.83
conf:(78.19)
5. SimDnkPlac_M SimNWork_M SingleParentF_H HomeOwn_L 31.01 ==> CAT_H 24.15
conf:(77.9)
6. SimBusi_M SimHSchool_M SingleParentF_H MUHouse_H 32.29 ==> CAT_H 25.08
conf:(77.66)

For medium-high CAT

7. Business_H SimMall_M HomeOwn_MH 32.42 ==> CAT_MH 25.99 conf:(80.16)
 8. Business_H SimWalmart_M SimMall_M HomeOwn_MH 30.61 ==> CAT_MH 24.29
conf:(79.36)
 9. SimPUnstableRent_M HomeOwn_MH MUHouse_M Income_L 30.25 ==> CAT_MH 23.96
conf:(79.2)
 10. SimDnkPlac_M HomeOwn_MH Employment_M UnstableRent_M 32.06 ==> CAT_MH
25.29 conf:(78.87)
-

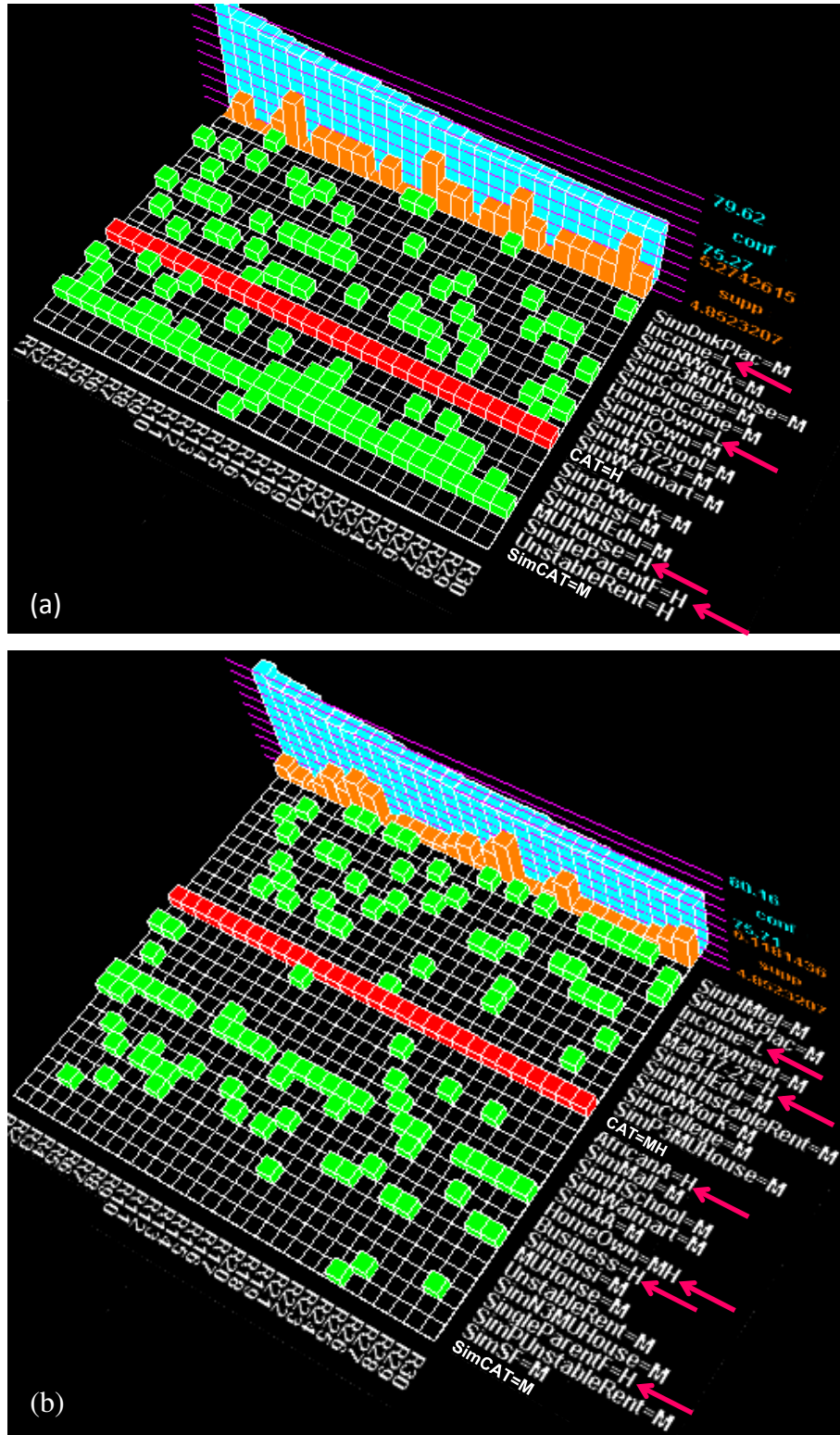


Figure 82: The strongest rules for (a) high and (b) medium high CAT

For MVT, the ten strongest rules mined by SpatialARMED are shown in Table 23 and the thirty strongest ones are visualized in Figure 83. A distinguishable association pattern is obtained for high MVT in this case, although at rather low degree of confidence (60-67%). The strongest rules indicate associations of unstable rent, low income, and heterogeneity to high MVT. On top of these factors, spillover effect, at a medium level, of multiple unit housing (Rule 1), or of Wal-Mart Super Center (R2), or of alcoholic drinking places (Rule 4) are also found in relation to high MVT at approximately 60% confidence. For areas which are heterogeneous and unstable rental but do not have low income, spillover effects of multiple-unit housing and business (Rule 7), or of shopping malls and young male (Rule 8), or of shopping malls and alcoholic drink places (Rule 10) are listed to be associated with high MVT.

Table 23: SpatialARMED strongest rules for MVT

For high MVT

1. SimP3MUHouse_M HeterogeneityInx_H UnstableRent_H Income_L 35.68 ==> MVT_H 23.95 conf:(67.12)
 2. SimWalmart_M HeterogeneityInx_H UnstableRent_H Income_L 37.07 ==> MVT_H 24.53 conf:(66.17)
 3. HeterogeneityInx_H UnstableRent_H Income_L 38.51 ==> MVT_H 24.96 conf:(64.81)
 4. SimDnkPlac_M HeterogeneityInx_H UnstableRent_H Income_L 38.04 ==> MVT_H 24.66 conf:(64.81)
 5. SimMall_M SimP3MUHouse_M HeterogeneityInx_H UnstableRent_H 38.77 ==> MVT_H 24.96 conf:(64.36)
 6. SimWalmart_M SimMall_M HeterogeneityInx_H UnstableRent_H 40.04 ==> MVT_H 25.37 conf:(63.36)
 7. SimBusi_M SimP3MUHouse_M HeterogeneityInx_H UnstableRent_H 39.74 ==> MVT_H 25.09 conf:(63.13)
 8. SimMall_M SimM1724_M HeterogeneityInx_H UnstableRent_H 38.02 ==> MVT_H 23.95 conf:(62.99)
 9. SimPUnstableRent_M SimP3MUHouse_M HeterogeneityInx_H UnstableRent_H 40.05 ==> MVT_H 25.23 conf:(62.98)
 10. SimDnkPlac_M SimMall_M HeterogeneityInx_H UnstableRent_H 40.57 ==> MVT_H 25.55 conf:(62.98)
-

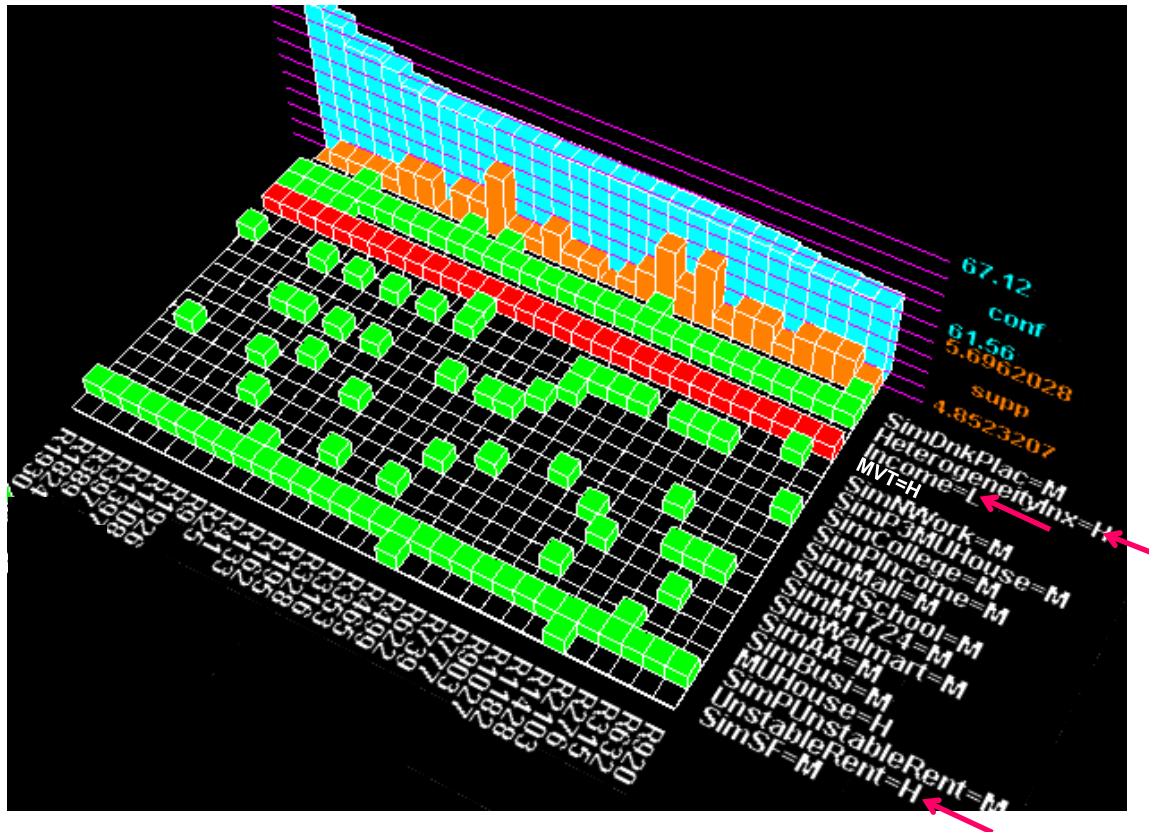


Figure 83: The strongest rules for high MVT

For TFM, the most frequent and strongest rules are listed in Table 24 and visualized in Figure 84. Although SpatialARMED suggested an identical list of known factors in comparison with that of CAT (see Table 21), examination on the strongest rules suggests different interactive patterns associating these factors to CAT and TFM. First of all, high TFM is also often associated with multiple-unit housing communities but at a much lower confidence level of 60-64%, rather than at 80% for CAT. For these multiple-unit housing communities, high concentration of African-Americans and spillover effect at medium level of ethnic heterogeneity are dominating associations to high TFM (Rule 1 to Rule 4). Spillover effects at medium level of hotels, motels or of business in addition are also found in these rules. The low confidence level at 60%, however, suggests that these

associations are not very strong. Second of all, medium-high TFM are often strongly (80-85%) found at neighborhoods of mix single family and multiple-unit housing due to the appearance of predicate “MUHouse_M” (Rule 5 to Rule 14). Particularly, the strongest rule in Table 24 indicates that if these areas of mix housing type are dominated by African-Americans, especially low income ones, and under spillover effect of heterogeneity at medium level then they experience medium-high TFM. High percentage of single-parent family also appears to associate with medium-high TFM, like for CAT. However, it is paired together with medium-high single family homeownership for FTM (Rule 11 and Rule 15), rather than with multiple-unit rental communities as previously shown in the case of CAT. Spillover impacts at medium level of crime attractors and generators such as alcoholic drinking places or Wal-Mart Super Center are also found together with high percentage of single-parent family and medium-high single family homeownership in relation to high TFM. Spillover effects at the medium level of heterogeneity appear very often in association to medium high TFM in general. Also, at 80% confidence, Rule 16 in Table 24 indicates spillover effect of nearby medium TFM, of nearby heterogeneity onto areas of some mix housing types and medium-high TFM.

Table 24: SpatialARMED strongest rules for TFM

For high TFM

1. SingleParentF_H MUHouse_H 37.55 ==> TFM_H 24.0 conf:(63.91)
2. SimHMtel_M SimP3MUHouse_M HomeOwn_L MUHouse_H 39.27 ==> TFM_H 25.01 conf:(63.67)
3. SimHMtel_M HomeOwn_L UnstableRent_H MUHouse_H 40.09 ==> TFM_H 25.19 conf:(62.82)
4. SimBusi_M SimHMtel_M HomeOwn_L MUHouse_H 39.76 ==> TFM_H 24.75 conf:(62.23)

For medium high TFM

5. SimHete_M AfricanA_H MUHouse_M Income_L 29.2 ==> TFM_MH 24.91 conf:(85.3)
 6. SimMall_M SimHete_M AfricanA_H MUHouse_M 28.93 ==> TFM_MH 24.54 conf:(84.8)
 7. SimHete_M AfricanA_H MUHouse_M 30.41 ==> TFM_MH 25.67 conf:(84.43)
 8. SimWalmart_M SimHete_M AfricanA_H MUHouse_M 29.37 ==> TFM_MH 24.65 conf:(83.95)
 9. SimWalmart_M SimHete_M SingleParentF_MH 29.06 ==> TFM_MH 24.04 conf:(82.71)
 10. SimHete_M SingleParentF_MH UnstableRent_M 30.55 ==> TFM_MH 25.12 conf:(82.22)
 11. SimWalmart_M SingleParentF_MH HomeOwn_MH HEdu_M 31.79 ==> TFM_MH 25.86 conf:(81.34)
 12. SimTFM_M SimHete_M HomeOwn_MH MUHouse_M 30.06 ==> TFM_MH 24.2 conf:(80.51)
 13. SingleParentF_MH HomeOwn_MH HEdu_M 33.75 ==> TFM_MH 27.13 conf:(80.39)
 14. SimCollege_M SimHete_M HomeOwn_MH MUHouse_M 30.76 ==> TFM_MH 24.7 conf:(80.3)
 15. SimDnkPlac_M SingleParentF_MH HomeOwn_MH HEdu_M 29.94 ==> TFM_MH 23.99 conf:(80.12)
-

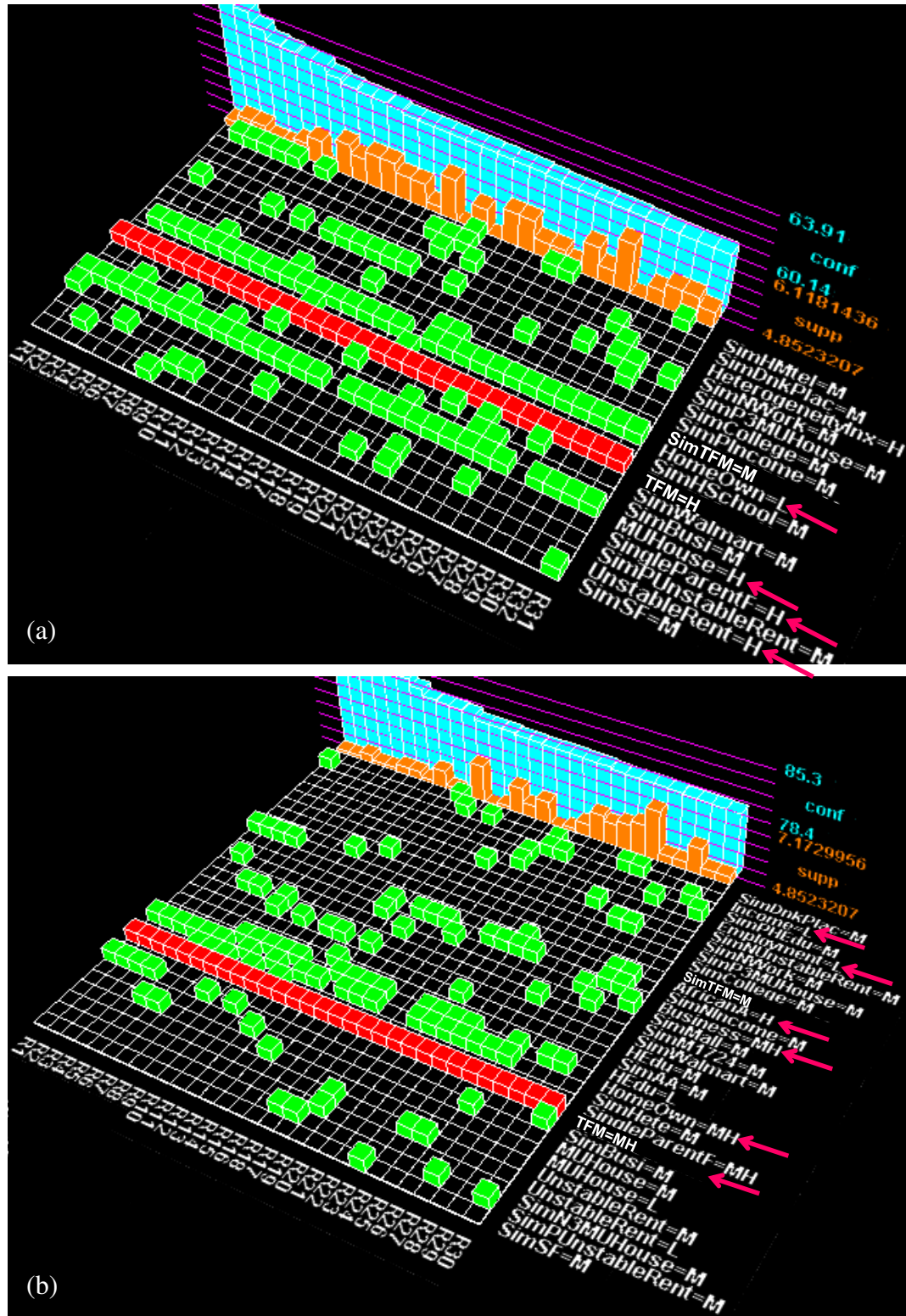


Figure 84: The strongest rules for high (a) and medium high (b) TFM

6.6.2.2 Examination based on Interactive Branching Approach

Examination and evaluation of the mined rules using the interactive branching approach for this case study further indicate a success of SpatialARMED in detecting interesting and potentially new associations to crime from the all perspectives of neighborhood effects, proximity effects, and interactive effects. Supporting evidence of this statement will be established through discussions in the next few paragraphs, subjected to strong (not exhaustive) findings.

The first step of rule evaluation following the iterative branching approach is to automatically extract the discovery predicates. Those associated to high and medium-high CAT, MVT, and TFM for this case study are reported in Table 25. These predicates, either one by one or in combinations, are detected with support larger than 5% for all types, confidence larger than 40% for MVT and 60% for CAT and TFM.

This set of discovering predicates demonstrates that the SpatialARMED framework has detected associations between high crime, particularly CAT and TFM, and the strong spillover impacts of known associations to crime such as multiple unit housing concentration, unemployment, young male concentration, business concentration, single parent family concentration, to name a few. The strong spillover impacts of crime itself are also detected in relation to high and medium-high CAT and TFM. For MVT, particularly, the set of discovering predicates shows that SpatialARMED only picked up medium-level spillover effects of MVT itself or of its associates on MVT. It should also be noted that the mined rules reported in this study are for particular thresholds of support (5%, equivalent to 23 group blocks out 474 ones) and confidence (40%). Considering that 5% support threshold is already low, one could argue for the use of an even lower one to

obtain a larger set of rules for MVT while hoping for more interesting findings, others dismiss the value of rules at such rare nature.

Table 25: List of discovering association predicates to crime with fuzzy mapping

Crime of all types		MVT	TFM	
SimCAT_H	HEdu_M	SimMVT_M	SimTFM_H	SimHete_M
SimCAT_M	SimHete_M	HEdu_M	SimTFM_M	MUHouse_M
SimDnkPlac_H	SimNIncome_M	SimP3MUHouse_M	SimDnkPlac_H	SimMall_M
SimHMtel_H	SimMall_M	SimPIncome_M	SimHMtel_H	SimWalmart_M
SimCollege_H	SimBusi_M	SimPUnstableRent_M	SimCollege_H	UnstableRent_M
SimAA_H	SimNWork_M	M	SimParkNRi_H	HEdu_M
SimPUnstableRent_H	SimWalmart_M	SimAA_M	SimHSchool_H	SimCollege_M
SimP3MUHouse_H	SimPUnstableRent_M	SimNHedu_M	SimAA_H	SimDnkPlac_M
SimN3MUHouse_H	MUHouse_M	SimPHedu_M	SimPUnstableRent_H	SimM1724_M
SimPIncome_H	SimDnkPlac_M	SimPWork_M	H	SimNWork_M
SimNIncome_H	Employment_M	SimM1724_M	SimP3MUHouse_H	SimNUnstableRent_M
SimNHedu_H	UnstableRent_M	SimSF_M	SimN3MUHouse_H	M
SimPWork_H	SimHSchool_M	SimHMtel_M	SimPIncome_H	SimPUnstableRent_M
SimNWork_H	SimNHedu_M	SimMall_M	SimNIncome_H	M
SimM1724_H	SimCollege_M	SimHSchool_M	SimHOwn_H	SimPHedu_M
SimSF_H	SimM1724_M	SimNWork_M	SimNHedu_H	SingleParentF_L
SimBusi_H	SimPWork_M	SimCollege_M	SimPWork_H	Employment_M
SimHete_H	SimPHedu_M	SimHOwn_M	SimNWork_H	SimHOwn_L
Business_MH	SimNUnstableRent_M	SimNUnstableRent_M	SimM1724_H	SimNUnstableRent_L
Business_L	M	SimDnkPlac_M	SimSF_H	L
Employment_H	SimHMtel_M	SimWalmart_M	SimBusi_H	SimParkNRi_L
HomeOwn_MH	SimN3MUHouse_M	SimBusi_M	Business_MH	SimN3MUHouse_L
SingleParentF_L	SimAA_M	SimNIncome_M	Business_L	SimBusi_L
AfricanA_L	SimP3MUHouse_M	SimHete_M	Employment_H	SimNHedu_L
AfricanA_M	SimPIncome_M		HomeOwn_MH	SimPWork_L
SimParkNRi_M	SimSF_M		HomeOwn_H	Income_M
	SimHOwn_M		MUHouse_L	SimHete_L
	MUHouse_L		Income_H	SimP3MUHouse_L
			HeterogeneityInx_L	SimNIncome_L
			UnstableRent_L	SimPIncome_L
			SimNUnstableRent_H	SimPUnstableRent_L
			Male17-24_L	SimSF_L
			AfricanA_L	SimM1724_L
			SimMall_L	SimHSchool_L
			SimN3MUHouse_M	SimPIncome_M
			SimAA_M	SimNHedu_M
			SimNIncome_M	SimPWork_M
			SimSF_M	SimHMtel_M
			SimP3MUHouse_M	SimHSchool_M
			SimBusi_M	SimParkNRi_M
				SimHOwn_M
				AfricanA_M

Among these discovering predicates, high home ownership (HomeOwn_H) and high income (Income_H) are reported in relation to high TFM. These associations are acknowledged in relation to TFM but with controversy as they are different from ones suggested by the social disorganization theory. In addition, it seems also interesting to see other associating factors, if exist, occurring with these in relation to high crime. They are thus chosen in this study for subgrouping rules.

Regarding high home ownership, medium-high home ownership has been detected previously among the strongest rules detected by SpatialARMED as an association, together with business concentration or with spillover effects of nearby neighborhoods which possess high rental multi-unit housing, low income, unstable rent and alcoholic drinking places, unemployed African-Americans, to high CAT. Here, it suggests that high home ownership, i.e. neighborhood dominated by owned properties, also associates with medium-high TFM. The ten strongest rules involving this relation are listed in Table 26 and the visualization of the first thirty ones is in Figure 85. At 69% confidence, SpatialARMED suggests that TFM activities in these high home ownership areas are strongly associated with highly employed, but low income or very heterogeneous population (Rule 2 and Rule 3). If low income and heterogeneity do not present but high employment, spillover effects at medium level of unemployment (Rule 4), or of high income and young male population (Rule 5), or of single parents and low education (Rule 7), or of nearby TFM (Rule 9, 10) are found associated to TFM in the areas of high home ownership.

Table 26: For areas with high homeownership: SpatialARMED strongest discovered associations

For medium-high TFM

1. HomeOwn_H Employment_H HEdu_M MUHouse_L 42.16 ==> TFM_MH 29.37 conf:(69.66)
2. HomeOwn_H HeterogeneityInx_H Employment_H 34.76 ==> TFM_MH 24.04 conf:(69.16)
3. HomeOwn_H MUHouse_L Income_L 36.32 ==> TFM_MH 24.67 conf:(67.91)
4. SimNWork_M HomeOwn_H Employment_H HEdu_M 36.55 ==> TFM_MH 24.7 conf:(67.57)
5. SimM1724_M SimPIncome_M HomeOwn_H Employment_H 35.62 ==> TFM_MH 23.99 conf:(67.35)
6. SimNIncome_M HomeOwn_H Employment_H MUHouse_L 38.51 ==> TFM_MH 25.87 conf:(67.17)
7. SimSF_M SimNHedu_M HomeOwn_H Employment_H 36.31 ==> TFM_MH 24.31 conf:(66.94)
8. SimWalmart_M HomeOwn_H Employment_H HEdu_M 45.25 ==> TFM_MH 30.27 conf:(66.9)
9. SimTFM_M SimNHedu_M HomeOwn_H Employment_H 36.46 ==> TFM_MH 24.38 conf:(66.86)
10. SimTFM_M SimNIncome_M HomeOwn_H Employment_H 37.59 ==> TFM_MH 25.05 conf:(66.65)

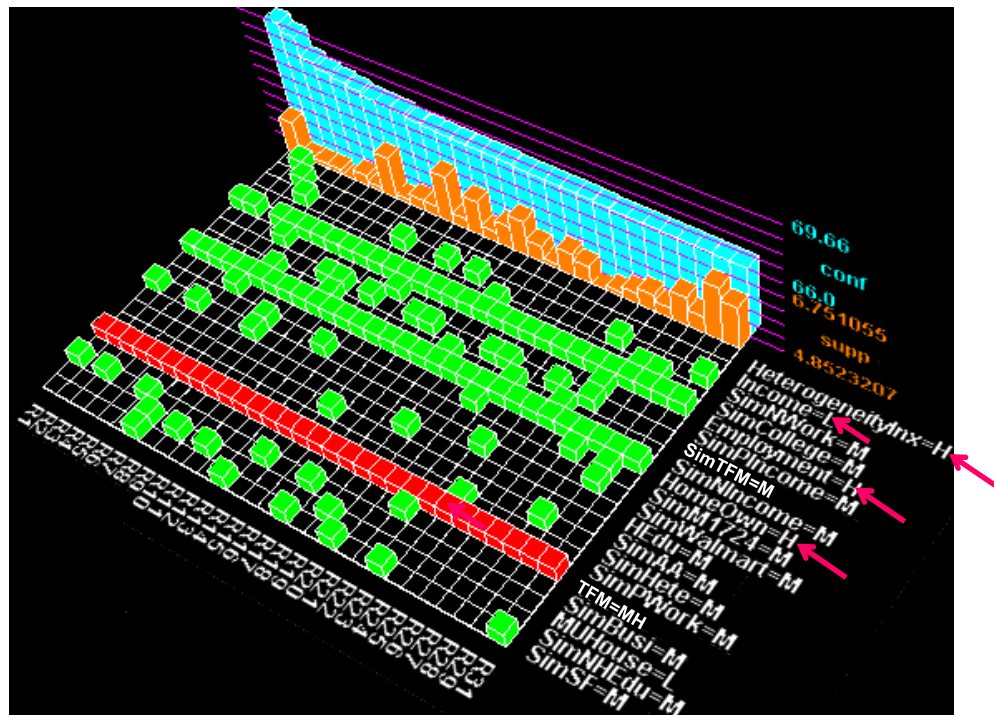


Figure 85: Potentially new fuzzy rules for TFM in association with high homeownership

Regarding high income, examination of the discovering predicate set reveals that areas of high income (presented by predicate “Income_H”) do not frequently experience

CAT, but medium-high TFM at support level of 5% and confidence level of 64%. Interestingly, the SpatialARMED mined rules suggest that neighborhoods located under strong spillover impact of high income (i.e. close to high income neighborhoods) also experience medium-high level of crime, but for both CAT and TFM and with very different association patterns.

Concerning high income and medium-high TFM, the ten strongest mined rules of this kind are examined and shown in Table 27. Their visualization is in Figure 86. By these rules, it is suggested that TFM in high income neighborhoods, at nearly 70% confidence, is related to either spillover effects of high schools, of low education, of Wal-Mart Super Center, or combinations among them. By extracting all the discovering predicates for rules of medium-high TFM and high income (shown in Table 28), it seems to be the case that the associations to TFM in rich neighborhoods are related to the spillover effects at medium level of nearby neighborhood effects and/or of nearby TFM.

Table 27: For areas has high income: strongest discovered SpatialARMED associations

For medium high TFM

1. SimHSchool_M SingleParentF_L HEdu_L Income_H 33.73 ==> TFM_MH 23.73 conf:(70.35)
2. SimHSchool_M HEdu_L Income_H 34.5 ==> TFM_MH 24.05 conf:(69.71)
3. SimWalmart_M SimHSchool_M HEdu_L Income_H 34.23 ==> TFM_MH 23.78 conf:(69.48)
4. SimHSchool_M AfricanA_L HEdu_L Income_H 34.23 ==> TFM_MH 23.78 conf:(69.47)
5. SingleParentF_L HEdu_L Income_H 39.21 ==> TFM_MH 27.06 conf:(69.01)
6. SingleParentF_L AfricanA_L HEdu_L Income_H 38.86 ==> TFM_MH 26.71 conf:(68.73)
7. SimWalmart_M SingleParentF_L HEdu_L Income_H 38.8 ==> TFM_MH 26.66 conf:(68.7)
8. HEdu_L Income_H 40.0 ==> TFM_MH 27.4 conf:(68.49)
9. AfricanA_L HEdu_L Income_H 39.65 ==> TFM_MH 27.05 conf:(68.22)
10. SimWalmart_M HEdu_L Income_H 39.59 ==> TFM_MH 27.0 conf:(68.18)

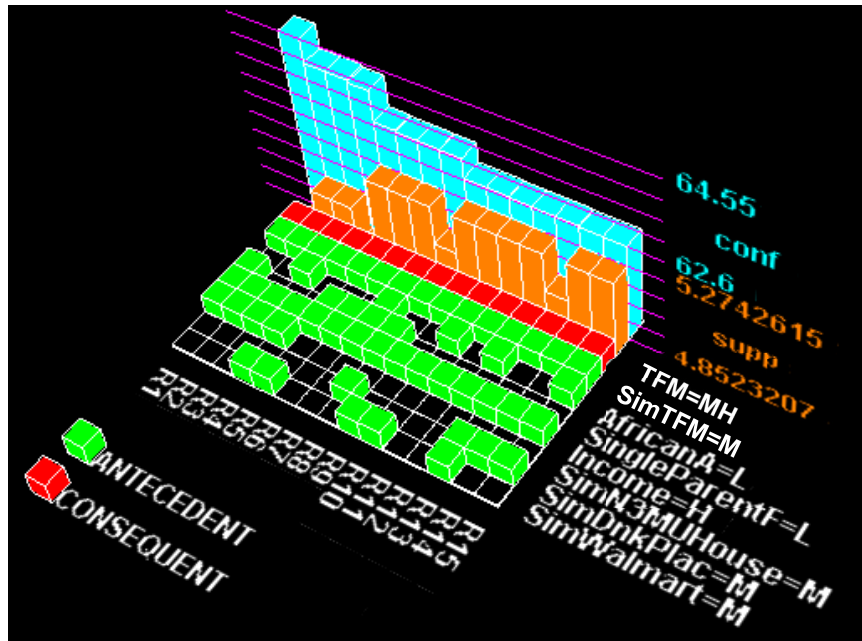


Figure 86: Fuzzy rules for TFM for areas has high income and medium spillover impact of TFM

Table 28: For areas that have high income: SpatialARMED associations

To Medium-high TFM:	
SingleParentF_L	SimTFM_M
HEdu_L	SimAA_M
AfricanA_L	SimSF_M
Income_H	SimHete_M
Male17-24_L	SimPWork_M
HeterogeneityInx_L	SimN3MUHouse_M
	SimNHedu_M
	SimBusi_M
	SimNWork_M
	SimP3MUHouse_M
	SimPHedu_M
	SimHSchool_M
	SimWalmart_M
	SimHMTel_M
	SimMall_M
	SimDnkPlac_M
	SimCollege_M
	SimParkNRi_M

Concerning CAT and TFM in neighborhoods near high income neighborhoods, lists of strongest rules are presented in Table 29 and Table 30, respectively, while their visualizations are in Figure 87 and Figure 88, respectively. Table 31 shows all the discovery predicates for rules of this kind. For CAT, it is suggested at a relatively high confidence level of near 70% that neighborhoods near high income, but also near unstable rent, multiple-unit housing , high young male population, and/or crime itself do experience medium-high CAT (Table 29 and Figure 87). At a rather lower confidence level of 65%, spillover effects of business concentration, of high employment, as well as of unemployment, are found associated to CAT in these neighborhoods. Associations to TFM in these neighborhoods however are found to be either the strong spillover effect of

nearly high schools (rules 3 and 4) or of nearby ethnic heterogeneity (rules 12 and 13), or of nearby business concentration (rules 14 and 15) (Table 30 and Figure 88). The strongest rules also suggest an association of nearby Wal-mart Super Centers at the medium level to TFM in neighborhoods near high income.

Table 30: For areas under strong influence of high income: SpatialARMED discovered associations

For medium high TFM

1. SimWalmart_M SimPIncome_H HEdu_L 41.79 ==> TFM_MH 25.95 conf:(62.11)
2. SimPIncome_H HEdu_L 42.25 ==> TFM_MH 25.98 conf:(61.5)
3. SimHSchool_M SimPIncome_H HeterogeneityInx_L 39.38 ==> TFM_MH 24.05 conf:(61.06)
4. SimWalmart_M SimHSchool_M SimPIncome_H HeterogeneityInx_L 39.38 ==> TFM_MH 24.05
conf:(61.06)
5. SimPIncome_H HeterogeneityInx_L 55.8 ==> TFM_MH 33.99 conf:(60.91)
6. SimWalmart_M SimPIncome_H HeterogeneityInx_L 55.8 ==> TFM_MH 33.99 conf:(60.91)
7. SimHOwn_H SimPIncome_H HeterogeneityInx_L 45.58 ==> TFM_MH 27.64 conf:(60.65)
8. SimWalmart_M SimHOwn_H SimPIncome_H HeterogeneityInx_L 45.58 ==> TFM_MH 27.64
conf:(60.65)
9. SimPHedu_M SimPIncome_H HeterogeneityInx_L 43.95 ==> TFM_MH 26.62 conf:(60.56)
10. SimWalmart_M SimPHedu_M SimPIncome_H HeterogeneityInx_L 43.95 ==> TFM_MH 26.62
conf:(60.56)
11. SimWalmart_M SimPIncome_H Male17-24_L 44.51 ==> TFM_MH 26.79 conf:(60.19)
12. SimHete_M SimPIncome_H HeterogeneityInx_L 45.8 ==> TFM_MH 27.54 conf:(60.13)
13. SimWalmart_M SimHete_M SimPIncome_H HeterogeneityInx_L 45.8 ==> TFM_MH 27.54 conf:(60.13)
14. SimBusi_H SimPIncome_H HeterogeneityInx_L 41.21 ==> TFM_MH 24.77 conf:(60.11)
15. SimBusi_H SimWalmart_M SimPIncome_H HeterogeneityInx_L 41.21 ==> TFM_MH 24.77
conf:(60.11)
16. SimPIncome_H Male17-24_L 44.62 ==> TFM_MH 26.8 conf:(60.07)

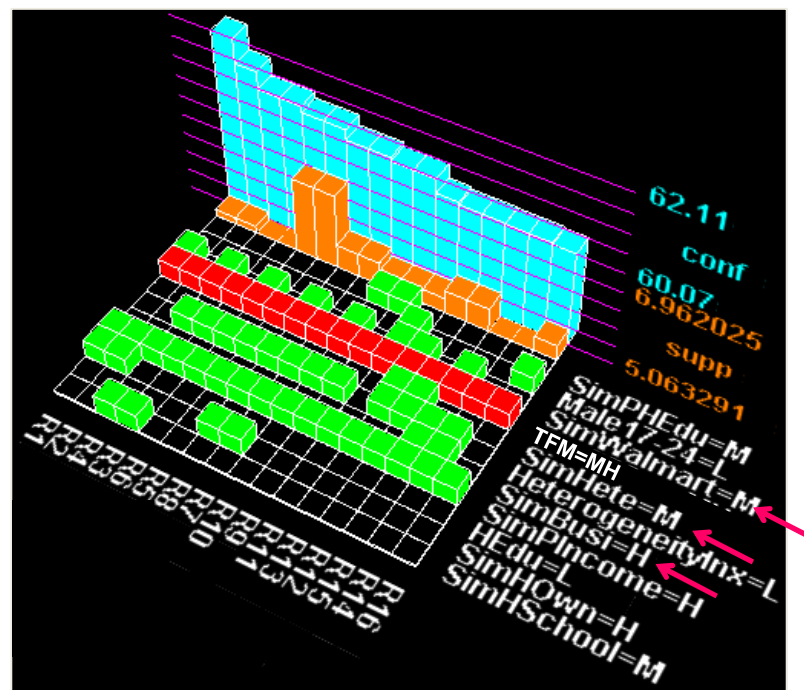


Figure 88: Fuzzy rules for TFM for areas under strong spillover impact high income

Table 31: For areas under strong influence of high income: SpatialARMED discovered associations

To Medium-high cime:	To Medium-high TFM:
SimCAT_H	HEdu_L
SimPUnstableRent_H	HeterogeneityInx_L
SimM1724_H	Male17-24_L
SimP3MUHouse_H	SimHOwn_H
SimPWork_H	SimBusi_H
SimBusi_H	SimHete_M
SimNWork_H	SimWalmart_M
SimNHedu_H	SimHSchool_M
SimWalmart_M	SimPHedu_M
SimParkNRi_M	
SimMall_M	

CHAPTER 7: CONCLUSIONS

This research has focused on the enlargement of association rule mining approaches for geographical knowledge discovery. Past studies in SAR mining have sidestepped the spatial aspects embedded within the geospatial problems at hand and overlooked the importance of interactive domain knowledge integration for rule evaluation. The overarching goal of this research thus has been to address these issues and to propose a comprehensive framework for spatial association rule mining and discovery. It has centered on the critical issues of how spatial concepts should be apprehended and how complex dependence structures, including spatial interactions, should be incorporated in both processes of mining and discovery while emphasizing visual analytics and domain knowledge integration for rule evaluation. A comprehensive framework, dubbed herein as SpatialARMED (Spatial Association Rule Mining and Discovery) has been proposed, which differs from earlier approaches of spatial association rule mining by capturing the spatial dependencies embedded in the data instead of relying on pre-conceived relationships. In addition, spatial interactions among associating factors to the phenomena under study are also modelled and taken into consideration. Composed of A Multi-directional Optimum Ecotope-Based Algorithm (AMOEBA) to detect spatial clusters, an operational spatial interactive model, a dependable mechanism for fuzzy mapping based predication, a popular association rule mining algorithm, and an effective domain knowledge integrated rule evaluation

procedure, the SpatialARMED framework presents an encouraging solution to extracting useful association patterns which enrich our understanding of spatial processes.

Another fundamental goal of this research is to validate the proposed framework and this has been achieved by applying SpatialARMED in criminology. The practice has not only demonstrated that SpatialARMED is practically implementable but it has also highlighted a number of critical contributions: First, particularly to criminology, frequent spatial association patterns to crime detected by SpatialARMED can be used to confirm the current knowledge of associations to crime as well as to further provide noteworthy insights in spatial processes of crime, including neighborhood effects, proximity effects and spatial/nonspatial interactions among participating associations to crime. Second, SpatialARMED is promising to outperform traditional spatial statistics on the capability to detect associations to the phenomenon under study. Traditional spatial statistics approaches work best with small scientifically sampled datasets under the limitations of confirmative hypotheses and assumptions of independent observations. They in addition adhere to limited perspectives, such as univariate spatial autocorrelation, or a specific and simple type of relation models, such as linear regression. SpatialARMED for SAR mining, on the other hand, is designed to handle multivariate spatial autocorrelation with complex, unpredictable non-linear relationships. Third, SpatialARMED offers the promise to surpass existing SAR mining algorithms by utilizing robust, defensible, and data-driven algorithms and procedures for quantifying spatial dependence structures, modelling the spillover impacts of these structures (i.e. spatial interactions among the participating variables), and integrating them into the process of mining. In comparison with existing SAR mining algorithms, SpatialARMED is robust in: (1) its definition of

HIGH or LOW defined on the basis of multi-directional searches for spatial clusters and tests for significance instead of pre-determined concept hierarchies; and (2) in its definition of NEAR-BY or NEXT-TO using domain knowledge based interaction models for spatial spillover impact.

From an application perspective, SpatialARMED presents a relevant data mining approach to extract spatial modalities governing the underlying processes observed in social sciences, regional studies, transportation, public health, business, marketing, human-environment interactions, and ecology, to name a few. In addition, SpatialARMED is designed to be scalable and to take advantage of high-performance computing for big-data analytics, which is in line with the future development of GIScience. By offering a mining solution which is (1) genuinely spatial-integrated, (2) resourceful for geographical big-data analysis, (3) an integrative platform for efficient powerful computing algorithms and domain expertise, this work contributes to the theoretical body of the literature in both entity-based spatial analysis and spatial data mining.

This research is not without limitations. First, the SpatialARMED framework encompasses several loosely-coupled algorithms and procedures which must be operated in a chronological sequence. While this offers implementation options to handle complexity and allows human interactions during the process of mining, it, indeed, is a limitation for applications that prioritize automation and timely responses. Second, the utilization of AMOEBA spatial clustering in this study detects a potential limitation tied to its difficulty to identify cold spots in cases where the data is highly skewed as discussed in Section 6.3.1. Further examination on this issue is given top priority in the

future research agenda. Finally, SpatialARMED at its current stage is limited to spatial analysis and need to be extended to spatial-temporal analysis. The research questions will then be expanded from mining spatial associations at one instance of time to mining spatial associations to changes over time or to analyze the changes in the spatial associations over time. This will be a fruitful research direction for the near future of SpatialARMED.

REFERENCES

- Abdi, H., and L.J. Williams. 2010. Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics* 2: 433–459.
- Adelman, R. M., T. Hui-shien, S. Tolnay, and K. D. Crowder. 2001. Neighborhood disadvantage among racial and ethnic groups: Residential location in 1970 and 1980. *The Sociological Quarterly* 42 (4): 603-632.
- Agarwal, P., and A. Skupin. 2008. *Self-organising maps: Applications in geographic information science*. Chichester: Wiley.
- Agarwal, R., C. Aggarwal, and V. Prasad. 2000. A tree projection algorithm for generation of frequent itemsets. *Parallel and Distributed Computing*.
- Agrawal, R. , and J. Shafer. 1996. Parallel mining of association rules. *IEEE Transactions on Knowledge and Data Engineering* 8 (6): 962-969.
- Agrawal, R., T. Imieliski, and A. Swami. 1993. "Mining association rules between sets of items in large databases." Presented at the The 1993 ACM SIGMOD international conference on Management of data, Washington, D.C., United States.
- Agrawal, R., and R. Srikant. 1995. "Mining sequential patterns." Presented at the Proceedings of the Eleventh International Conference on Data Engineering.
- Agrawal, R., and R. Srikant, eds. 1994. *Fast algorithms for mining association rules* Edited by J. B.; Jarke Bocca, M.; Zaniolo, C. . Vol. 1215, *Organization*.
- Aldstadt, J., and A. Getis. 2006. Using amoeba to create a spatial weights matrix and identify spatial clusters. *Geographical Analysis* 38: 327-343.
- Andresen, M.A. 2006. Crime measures and the spatial analysis of criminal activity *British Journal of Criminology* 46 (2): 258-28.
- Anselin, L. 1986a. Non-nested tests on the weight structure in spatial autoregressive models: Some monte carlo results. *Journal of Regional Science* 26 (2).
- Anselin, L. 1986b. Some further notes on spatial models and regional science. *Journal of Regional Science* 26 (4): 799-802.
- Anselin, L. 1995. Local indicators of spatial association - lisa. *Geographical Analysis* 27: 93–115.
- Anselin, L. 2003. Spatial externalities, spatial multipliers, and spatial econometrics. *International Regional Science Review* 26: 153–166.

- Anselin, L., ed. 1999. *Interactive techniques and exploratory spatial data analysis*. Edited by A.P. Longley, M.F. Goodchild, D.J. Maguire and D.W. Rhind, *Geographical information systems—principles and technical issues*. New York, NY: John Wiley & Sons, Inc.
- Anselin, L. . 1988. *Spatial econometrics: Methods and models*. Dordrecht: Kluwer.
- Anselin, L. . 1990. Spatial dependence and spatial structural instability in applied regression analysis. *Journal of Regional Science* 30 (2): 185-207.
- Anselin, L., J. Cohen, D. Cook, W. Gorr, and Tita G. 2000. Spatial analysis of crime. *Measurement and Analysis of Crime and Justice* 2000 4: 213-262.
- Anselin, L., and D. A. Griffith. 1988. Do spatial effects really matter in regression analysis? *Papers of the Regional Science Association* 65: 11-34.
- Appice, A, M. Ceci, A. Lanza, F.A. Lisi, and D. Malerba. 2003. Discovery of spatial association rules in geo-referenced census data: A relational mining approach. *Intelligent Data Analysis* 7: 541-566.
- Appice, A., M. Berardi, M. Ceci, and D. Malerba, eds. 2005. *Mining and filtering multi-level spatial association rules with ares*. Edited by M-S. Hacid, N. Murray, Z. Ras and S. Tsumoto. Vol. 3488, *Foundations of intelligent systems - lecture notes in computer science*: Springer Berlin / Heidelberg.
- Asanovic, K., R. Bodik, B.C. Catanzaro, J.J. Gebis, and P. Husbands. 2006. *The landscape of parallel computing research: A view from berkeley*. University of California, Berkeley. Technical Report No. UCB/EECS-2006-183.
- Ashrafi, M. , D. Taniar, and K. Smith. 2005. Redundant association rules reduction techniques. *Lecture Notes in Computer Science* 3809: 254 - 263.
- Ashrafi, M., D. Taniar, and K. Smith. 2004. A new approach of eliminating redundant association rules. *Lecture Notes in Computer Science* 3180: 465-474.
- Bailey, T. C., and A. C. Gatrell. 1995. *Interactive spatial data analysis*. New York, NY: John Wiley and Sons, Inc.
- Baller, R.D., L. Anselin, S.F. Messner, G. Deane, and D.F. Hawkins. 2001. Structural covariates of u.S. County homicide rates: Incorporating spatial effects. *Criminology* 39: 561–590.
- Baralis, E., and G. Psaila. 1997. Designing templates for mining association rules. *Journal of Intelligent Information Systems* 9 (1): 7-32.
- Barney, B. 2013. "Introduction to parallel computing." Jul 15, 2013 (Nov 5, 2013 2013).
- Benzecri, J.-P. 1973. *L'analyse des donnees*. Paris Dunod

- Bernasco, W., and R. Block. 2011. Robberies in Chicago: A block-level analysis of the influence of crime generators, crime attractors, and offender anchor points. *Journal of Research in Crime and Delinquency* 48 (1): 33-57.
- Bernasco, W., and H. Elffers. 2010. "Statistical analysis of spatial crime data." In *Handbook of quantitative criminology*, eds. Alex R. Piquero and David Weisburd. New York: Springer. 699-724.
- Bertino, E., and M. L. Damiani. 2005. Spatial knowledge-based applications and technologies: Research issues. *Knowledge-Based Intelligent Information and Engineering Systems, Pt 4, Proceedings* 3684: 324-328.
- Birks, D., M. Townsley, and A. Stewart. 2012. Generative explanations of crime: Using simulation to test criminological theory. *Criminology* 50 (1): 221-254.
- Bogges, L. N., and J. R. Hipp. 2010. Violent crime, residential instability and mobility: Does the relationship differ in minority neighborhoods? *Journal of Quantitative Criminology* 26: 351-370.
- Bogorny, V. 2006a. "Enhancing spatial association rule mining in geographic databases." Ph.D. Universidade Federal DO RIO grande Do Sul.
- Bogorny, V., S. Camargo, P.M. Engel, and L. O. Alvares. 2006a. "Mining frequent geographic patterns with knowledge constraints." In *ACM-GIS'06*. Arlington, Virginia, USA.
- Bogorny, V., Sd S. Camargo, P. M. Engel, and L. O. Alvares. 2006b. "Towards elimination of well known geographic patterns in spatial association rule mining." In *the 3rd International IEEE Conference on Intelligent Systems*. 532-537.
- Bogorny, V., P.M. Engel, and L.O. Alvares. 2005. "Towards the reduction of spatial join for knowledge discovery in geographic databases using geo-ontologies and spatial integrity constraints." In *Workshop on knowledge discovery and ontologies of the ECML/PKDD*.
- Bogorny, V., P.M. Engel, and L.O. Alvares, eds. 2008a. *Enhancing spatial association rule mining in geographic databases using geo-ontologies*. Edited by H.O. NIGRO, S.G. Cisaró and S.G. Xodo, *Data mining with ontologies: Implementations, findings and frameworks*: Idea Group Inc.
- Bogorny, V., B. Kuijpers, and L. O. Alvares. 2008b. Reducing uninteresting spatial association rules in geographic databases using background knowledge: A summary of results. *International Journal of Geographical Information Science* 22 (4): 361-386.
- Bogorny, V., J. Valiati, and L. Alvares. 2010. Semantic-based pruning of redundant and uninteresting frequent geographic patterns. *GeoInformatica* 14 (2): 201-220.

- Bogorny, V.; Engel, P.M.; Alvares, L.O. 2006b. "Geoarm: An interoperable framework to improve geographic data preprocessing and spatial association rule mining." In *The 18th conference on software engineering and knowledge engineering (SEKE'06)*. San Francisco, California. 79-84.
- Bogorny, V.; Palma, A.T.; Engel P.M.; Alvares L.O. 2006c. "Weka-gdpm: Integrating classical data mining toolkit to geographic information systems." In *SBBT Workshop on Data Mining Algorithms and Applications (WAAMD'06)*. Florianopolis, Brazil. 9-16.
- Bosse, T., H. Elffers, and Gerritsen. C. 2010. Simulating the dynamical interaction of offenders, targets and guardians. *Crime Patterns and Analysis* 3: 51-66.
- Bottoms, A. E., and P. Wiles, eds. 2002. *Environmental criminology*. Edited by M. Maguire, R. Morgan and R. Reiner, *The oxford handbook of criminology (3th edition)*. Oxford: Oxford University Press.
- Bouckaert, R. R., E. Frank, M. Hall, R. Kirkby, P. Reutemann, A. Seewald, and D. Scuse. 2011. *Weka manual for version 3-7-5*. The University of Waikato.
- Brantingham, P. , and P. Brantingham. 1995. Criminality of place: Crime generators and crime attractors. *European Journal of Criminal Policy and Research* 3: 5-26.
- Brantingham, P., and P. Brantingham. 1981. *Environmental criminology*. Sage Publications.
- Brantingham, P., and P. Brantingham. 1984. *Patterns in crime*. New York: Macmillan.
- Brantingham, P. J., and P. L. Brantingham, eds. 2008. *Crime pattern theory*. Edited by R. Wortley and L. Mazerolle, *Environmental criminology and crime analysis*. Portland, OR: Willan Publishing.
- Brantingham, P. J., and G. E. Tita, eds. 2008. *Offender mobility and crime pattern formation from first principles*. Edited by L. Liu and J.E. Eck, *Artificial crime analysis systems: Using computer simulations and geographic information systems*. Hershey, PA: Idea Group.
- Brantingham, P.J. 2013. Prey selection among los angeles car thieves. *Crime Science* 2 (3).
- Brin, S., R. Motwani, and C. Silverstein. 1997a. "Beyond market baskets: Generalizing association rules to correlations." In *ACM SIGMOD International Conference on Management of Data*. 265-276.
- Brin, S., R. Motwani, J.D. Ullman, and Tsur S. 1997b. Dynamic itemset counting and implication rules for market basket data. *ACM Press, SIGMOD Record* 6 (2): 255-264.

- Brooks-Gunn, J., G. J. Duncan, and J. L. Aber, eds. 1997. *Neighborhood poverty: Vol i: Context and consequences for children*. New York: Russell Sage.
- Brunsdon, C., A.S. Fotheringham, and M.E. Charlton. 1996. Geographically weighted regression: A method for exploring spatial nonstationarity. *Geographical Analysis* 28 (4): 281–298.
- Bruzzese, D., and C. Davino. 2008. "Visual mining of association rules." In *Visual data mining*, eds. S.J. Simoff, M.H.B. Hlen and A. Mazeika: Springer-Verlag. 103-122.
- Bruzzese, D., C. Davino, and D. Vistocco. 2003. "Parallel coordinates for interactive exploration of association rules." In *Proceedings of the 10th International Conference on Human - Computer Interaction*. Creta, Greece: Lawrence Erlbaum, Mahwah
- Burgess, E. W., ed. 1925. *The growth of the city*. Edited by R.E. Park, E.W. Burgess and R.D. McKenzie, *The city chicago*: University of Chicago Press.
- Bursik, R. J., and H. G. Grasmick. 1993. *Neighborhoods and crime: The dimensions of effective community control*. New York: Lexington Books.
- Cahill, M. , and G. Mulligan. 2007. Using geographically weighted regression to explore local crime patterns. *Social Science Computer Review* 25: 174-193.
- Ceri, S., F. Cacace, and L. Tanca. 1991. Object orientation and logic programming for databases - a seasons flirt or long-term marriage. *Lecture Notes in Computer Science* 504: 124-143.
- Chainey, S., and J. Ratcliffe. 2005. *Gis and crime mapping*. London: John Wiley & Sons.
- Chainey, S., S. Reid, and N. Stuart. 2002. "When is a hotspot a hotspot? A procedure for creating statistically robust hotspot maps of crime." In *Innovations in gis 9: Socio-economic applications of geographic information science*, eds. D. Kidner, G. Higgs and S. White. London: Taylor and Francis. 21-36.
- Chen, H., W. Chung, J.J. Xu, G. Wang, Y. Qin, and M. Chau. 2004. Crime data mining: A general framework and some examples. *Computer* 37 (4): 50-56.
- Chernoff, H., and M. H. Rizvi. 1975. Effect on classification error of random permutations of features in representing multivariate data by faces. *Journal of American Statistical Association* 70: 548–554.
- Cheung, D., and Y. Xiao. 1998. Effect of data skewness in parallel mining of association rules. *Lecture Notes in Computer Science* 1394: 48-60.
- Cheung, D.W.L., J. Han, V.T. Ng, A.W. Fu, and Y. Fu. 1996. "A fast distributed algorithm for mining association rules." In *International Conference on Parallel and Distributed Systems*. 31-42.

- Chuang, K-T., M-S. Chen, and W-C. Yang. 2005. "Progressive sampling for association rules based on sampling error estimation advances in knowledge discovery and data mining." In, eds. T. Ho, D. Cheung and H. Liu: Springer Berlin / Heidelberg. 37-44.
- Clarke, R. V., and M. Felson, eds. 1993. *Introduction: Criminology, routine activity, and rational choice*. Edited by R. V. Clarke and M. Felson, *Routine activity and rational choice*. New Brunswick: Transaction publishers.
- Clarke, R. V., and P. M. Harris. 1992. *Auto theft and its prevention*. Chicago: University of Chicago Press.
- Clarke, R.V. 2010. "Theft of and from cars in parking facilities (problem-oriented guides for police series no. 10.) ": Washington, DC: Office of Community Oriented Policing Services, U.S. Department of Justice.
- Cliff, A. D. , and J. K. Ord, eds. 1969. *The problem of spatial autocorrelation*. Edited by A. J. Scott, *London papers in regional science 1, studies in regional science*. London: Pion.
- Cliff, A. D., and J. K. . Ord. 1973. *Spatial autocorrelation*. London: Pion.
- Cliff, A. D., and K. Ord. 1970. Spatial autocorrelation: A review of existing and new measures with applications. *Economic Geography* 46 (ArticleType: research-article / Issue Title: Supplement: Proceedings. International Geographical Union. Commission on Quantitative Methods / Full publication date: Jun., 1970 / Copyright © 1970 Clark University): 269-292.
- Coenen, F. 2004. The lucs-kdd apriori-t association rule mining algorithm. *Department of Computer Science, The University of Liverpool*.
- Cohen, L., and M. Felson. 1979. Social change and crime rate trends: A routine activity approach. *American Sociological Review* 44: 488-608.
- Cook, D., A. Buja, J. Cabrera, and C. Hurley. 1995. Grand tour and projection pursuit. *Journal of Computational and Graphical Statistics* 4 (3): 155–172.
- Copes, H. 1999. Routine activities and motor vehicle theft: A crime specific approach. *Journal of Crime and Justice* 22 (2): 125–146.
- Copes, H. 2003. Streetlife and the rewards of auto theft. *Deviant Behavior* 24: 309-332.
- Cornish, D., and R. Clarke. 1986. *The reasoning criminal: Rational choice perspectives on offending*. New York: Springer-Verlag.
- Cristofor, L., and D. Simovici. 2002. "Generating an informative cover for association rules." In *The IEEE International Conference on Data Mining*.

- Dacey, M.F., ed. 1968. *A review of measures of contiguity for two- and k-color maps*. Edited by B. J. L. Berry and D. F. Marble, *Spatial analysis: A reader in statistical geography*. Englewood Cliffs, N.J.: Prentice-Hall.
- Davison, E.L. 1995. "An ecological analysis of crime in a mid-sized southern city: Tests of routine activity and social disorganization approaches." Unpublished doctoral dissertation, North Carolina State University, Raleigh.
- Deane, G., S. Messner, T. Stucky, K. McGeever, and C. Kubrin. 2008. Not 'islands, entire of themselves': Exploring the spatial context of city-level robbery rates. *Journal of Quantitative Criminology* 24: 337–421.
- Do, T., S. Hui, and A. Fong. 2003. "Mining frequent itemsets with category-based constraints discovery science." In, eds. G. Grieser, Y. Tanaka and A. Yamamoto: Springer Berlin / Heidelberg. 76-86.
- Dray, A., L. G. Mazerolle, P. Perez, and A. Ritter, eds. 2008. *Drug law enforcement in an agent-based model: Simulating the disruption*. Edited by L. Liu and J. E. Eck, *Artificial crime analysis systems: Using computer simulations and geographic information systems*. Hershey, PA: Idea Group.
- Duque, J.C., B. Dev, A. Betancourt, and J.L. Franco. 2011. "Clusterpy: Library of spatially constrained clustering algorithms, version 0.9.9." In: RiSE-group (Research in Spatial Economics). EAFIT University.
- Dykes, J. 1998. Cartographic visualization: Exploratory spatial data analysis with local indicators of spatial association using tcl/tk and cdv. *The Statistician* 47 (3): 485-497.
- Eck, J. , S. Chainey, J. G. James G. Cameron, M. Leitner, and R. E. Wilson 2005. *Mapping crime: Understanding hot spots* Washington, DC: National Institute of Justice.
- Egenhofer, M.J., and R.D. Franzosa. 1991. Point-set topological spatial relations. *International Journal of Geographical Information Systems* 5 (2): 161-174.
- Epstein, J.M., and R. Axtell. 1996. *Growing artificial societies: Social science from the bottom up*. Brookings Institution Press.
- Ester, M., H.-P. Kriegel, and J. Sander, eds. 2001. *Algorithms and applications for spatial data mining*. Edited by M. Harvey and J. Han, *Geographic data mining and knowledge discovery, research monographs in gis*: Taylor and Francis.
- Estivill-Castro, V., and I. Lee. 2001. "Data mining techniques for autonomous exploration of large volumes of geo-referenced crime data." In *The 6th International Conference in Geocomputation*. Brisbane, Australia.
- Fayyad, U., G. Piatetsky-Shapiro, and P. Smyth, eds. 1996. *From data mining to knowledge discovery—an review*. Edited by U. Fayyad, G. Piatetsky-Shapiro, P. Smyth

- and R. Uthurusay, *Advances in knowledge discovery*. Cambridge, MA: AAAI Press/The MIT Press.
- Felson, M., ed. 1995. *Those who discourage crime*. Edited by D. Weisburd and J. Eck, *Crime and place*. Monsey, NY: Criminal Justice Press.
- Florax, R., and S. Rey, eds. 1995. *The impacts of misspecified spatial interaction in linear regression models*. Edited by L. Anselin and R.J.G.M. Florax, *New directions in spatial econometrics*. Berlin: Springer.
- Folmer, H., and J. Oud. 2008. How to get rid of w: A latent variables approach to modelling spatially lagged variables. *Environment and Planning A* 40: 2526-2538.
- Fotheringham, A. S. , and D. W. S. Wong. 1991. The modifiable areal unit problem in multivariate statistical analysis. *Environment and Planning A* 23: 1025-1044.
- Fotheringham, A.S., C. Brunsdon, and M. Charlton. 2002. *Geographically weighted regression: The analysis of spatially varying relationships*. West Sussex, UK: Wiley.
- Geary, R. C. 1954. The contiguity ratio and statistical mapping. *The Incorporated Statistician* 5 (3): 115-146.
- Getis, A. 2008. A history of the concept of spatial autocorrelation: A geographer's perspective. *Geographical Analysis* 40 (3): 297-309.
- Getis, A. 2009. Spatial weights matrices. *Geographical Analysis* 41 (4): 404-410.
- Getis, A., and J. Aldstadt. 2004. Constructing the spatial weights matrix using a local statistic. *Geographical Analysis* 36 (2): 90-104.
- Getis, A., and J. K. Ord. 1992. The analysis of spatial association by use of distance statistics. *Geographical Analysis* 24: 189-206.
- Greenacre, M. 1993. *Correspondence analysis in practice*. London: Academic Press.
- Griffith, D.A., ed. 1996. *Some guidelines for specifying the geographic weights matrix contained in spatial statistics models*. Edited by S.L. Arlinghaus, D.A. Griffith, W.C. Arlinghaus, W.D. Drake and J.D. Nystrom, *Practical handbook of spatial analysis*. FL, London and New York: CRC Press.
- Groff, E. R. 2007. Simulation for theory testing and experimentation: An example using routine activity theory and street robbery. *Journal of Quantitative Criminology* 25: 75-103.
- Groff, E. R. 2008. Adding the temporal and spatial aspects of routine activities: A further test of routine activity theory. *Security Journal* 21: 95-116.

- Guo, D., M. Gahegan, A. M. MacEachren, and B. Zhou. 2005. Multivariate analysis and geovisualization with an integrated geographic knowledge discovery approach. *Cartography and Geographic Information Science* 32 (2): 113-132.
- Guo, D., D. Peuquet, and M. Gahegan. 2003. Iceage: Interactive clustering and exploration of large and high-dimensional geodata. *Geoinformatica* 7 (3): 229-253.
- Guttman, A. 1984. "R-trees a dynamic index structure for spatial searching." In *INTERNATIONAL CONFERENCE ON MANAGEMENT OF DATA*. Boston, Massachusetts.
- Haining, R. 1977. Model specification in stationary random fields. *Geographical Analysis* 9: 107-129.
- Haining, R. 1979. Statistical tests and process generators for random field models. *Geographical Analysis* 11: 45-64.
- Haining, R. 1986. Spatial models and regional science: A comment on anselin's paper and research directions. *Journal of Regional Science* 26 (4): 793-798.
- Haining, R. 2003. *Spatial data analysis: Theory and practice*. Cambridge: University Press.
- Haining, R. P. 1978. Estimating spatial-interaction models. *Environment and Planning A* 10 (3): 305-320.
- Han, E-H., G. Karypis, and V. Kumar. 1997a. " Scalable parallel data mining for association rules." In *ACM SIGMOD conference on management of data*.
- Han, J., and M. Kamber. 2001. *Data mining concepts and techniques*. Morgan Kaufmann.
- Han, J., M. Kamber, and A. K. H. Tung, eds. 2001. *Spatial clustering methods in data mining: A survey*. Edited by H. J. Miller and J. Han, *Geographic data mining and knowledge discovery*. London: Taylor and Francis.
- Han, J., K. Koperski, and N. Stefanovic. 1997b. "Geominer: A system prototype for spatial data mining." In.
- Han, J., and J. Pei. 2000. Mining frequent patterns by pattern-growth: Methodology and implications. *ACM SIGKDD Explorations Newsletter* 2 (2): 14-20.
- Harris, R., J. Moffat, and V. Kravtsova. 2011. In search of 'w'. *Spatial Economic Analysis* 6 (3): 249-270.
- Hetzler, B., W. M. Harris, S. Havre, and P. Whitney. 1998. "Visualizing the full spectrum of document relationships." In *The Fifth International Society for Knowledge Organization (ISKO) Conference*.

- Hipp, J.R. 2007. Income inequality, race, and place: Does the distribution of race and class within neighborhoods affect crime rates? *Criminology* 45 (3): 665-695.
- Hofmann, H., A. Siebes, and A. Wilhelm. 2000a. "Visualizing association rules with interactive mosaic plots." In *The 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 227–235.
- Hofmann, H., A.P.J.M. Siebes, and A.F.X. Wilhelm. 2000b. "Visualizing association rules with interactive mosaic plots." In *the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*: ACM Press. 227–235.
- Hofmann, H., and A. Wilhelm. 2001. Visual comparison of association rules. *Computational STATISTICS IN MEDICINE* 16: 399–416.
- Hubert, L. 1985. Combinatorial data analysis: Association and partial association. *Psychometrika* 50: 449-467.
- Hubert, L. J., R. G. Golledge, and C. M. Costanza. 1981. Generalized procedures for evaluating spatial autocorrelation. *Geographical Analysis* 13: 224–232.
- Jacox, E. , and H. Samet. 2007. Spatial join techniques. . *ACM Transactions on Database Systems* 32 (1).
- Jaroszewicz, S., and D. Simovici. 2002. Pruning redundant association rules using maximum entropy principle. *Lecture Notes in Computer Science* 2336: 135-142.
- Jefferis, E. 1999. "A multi-method exploration of crime hot-spots: A summary of findings." National Institute of Justice.
- Jung, C., and C-H. Sun. 2006. "Development of a giservice based on spatial data mining for location choice of convenience stores in taipei city." In, eds. H. Wu and Q. Zhu. 642117-642117-10.
- Keim, D., F. Mansmann, J. Schneidewind, J. Thomas, and H. Ziegler. 2008. "Visual analytics: Scope and challenges." In *Visual data mining: Theory, techniques and tools for visual analytics*, eds. S.J. Simoff, M.H. Böhlen and A. Mazeika: Springer Berlin / Heidelberg. 76-90.
- Keister, T. 2007. "Thefts of and from cars on residential streets and driveways (problem-oriented guides for police series no. 46)." Washington, DC: Office of Community Oriented Policing Services, U.S. Department of Justice.
- Keyvanpour, M. R., and M. R. Ebrahimi. 2011. Detecting and investigating crime by means of data mining: A general crime matching framework. *Procedia Computer Science* 3 (0): 872-880.

- Kim, Y., and N. Xiao, eds. 2008. *Fraudsims: Simulating fraud in a public delivery program*. Edited by L. Liu and J. E. Eck, *Artificial crime analysis systems: Using computer simulations and geographic information systems*. Hershey, PA: IGI Global.
- Klir, G.J., U.St. Clair, and B. Yuan. 1997. *Fuzzy set theory: Foundations and applications*. Prentice Hall.
- Klosgen, W., and M. May. 2002. "Spatial subgroup mining integrated in an object-relational spatial database." Presented at the Proceedings of the 6th European Conference on Principles of Data Mining and Knowledge Discovery.
- Koperski, K., J. Adhikary, and J. Han. 1996. "Knowledge discovery in spatial databases: Progress and challenges " In *ACM SIGMOD workshop on research issues on data mining and knowledge discovery (DMKD'96)*. Montréal: IRIS/Precarn. 55-70.
- Koperski, K., and J. Han. 1995. "Discovery of spatial association rules in geographic information databases." In *Advances in spatial databases*, eds. M.J. Egenhofer and J.R. Herring: Springer Berlin / Heidelberg. 47-66.
- Koperski, K., J. Han, and N. Stefanovic. 1998. "An efficient two-step method for classification of spatial data." In.
- Kornhauser, R. 1987. *Social sources of delinquency*. Chicago: University of Chicago Press.
- Kostov, P. 2010. Model boosting for spatial weight matrix selection in spatial lag model. *Environment and Planning B: Planning and Design* 37: 533-549.
- Kotsiantis, S., and D. Kanellopoulos. 2006. Association rules mining: A recent overview. *International Transactions on Computer Science and Engineering* 32 (1): 71-82.
- Kposowa, A., and K.D. Breault. 1993. Reassessing the structural covariates for u.S. Homicide rates: A county level study. *Sociological Forces* 26: 27-46.
- Krivo, L.J., and R.D. Peterson. 1996. Extremely disadvantaged neighborhoods and urban crime. *Social Forces* 75: 619-650.
- Kubrin, C.E. 2003. Structural covariates of homicide rates: Does type of homicide matter? *Journal of Research in Crime and Delinquency* 40: 139-170.
- Ladner, R., F. E. Petry, and M. A. Cobb. 2003. Fuzzy set approaches to spatial data mining of association rules. *Transactions in GIS* 7 (1): 123-138.
- Land, K., P. McCall, and L. Cohen. 1990. Structural covariates of homicide rates: Are there invariances across time and social space? *American Journal of Sociology* 95: 922-963.

- Laney, D. 2001. 3d data management: Controlling data volume, velocity, and variety. *Application delivery strategies*, META Group Inc.
- Laney, D. 2012. *The importance of 'big data': A definition*. Gartner.
- LeBeau, J. L. 1987. The methods and measures of centrography and the spatial dynamics of rape. *Journal of Quantitative Criminology* 3 (2): 125-141.
- LeBeau, J. L. 1992. Four case studies illustrating the spatial-temporal analysis of serial rapists. *Police Studies* 15 (3): 124-145.
- Lee, I., and V. Estivill-Castro. 2006. Fast cluster polygonization and its applications in data-rich environments. *Geoinformatica* 10: 399-422.
- Lee, I., and V. Estivill-Castro. 2011. Exploation of massive crime data sets through data mining techniques. *Applied Artificial Intelligence* 25 (5): 362-379.
- Lee, I., and P. Phillips. 2008. Urban crime analysis through areal categorized multivariate associations mining. *Applied Artificial Intelligence* 22 (5): 483-499.
- Lee, M., M. Maume, and G. Ousey. 2003 Social isolation and lethal violence across the metro/nonmetro divide: The effects of socioeconomic disadvantage and poverty concentration on homicide. . *Rural Sociology* 68: 107-131.
- LeSage, J.P., ed. 2004. *A family of geographically weighted regression models*. Edited by L. Anselin, R.J.G.M. Florax and S.J. Rey, *Advances in spatial econometrics: Methodology, tools and applications*. Berlin: Springer.
- LeSage, J.P., and R. K. Pace. 2010. "The biggest myth in spatial econometrics."
- Li, X. 2008. "Mining spatial association rules in spatially heterogeneous environment." In *International Conference on Earth Observation Data Processing and Analysis (ICEODPA)*, eds. D. Li, J. Gong and H. Wu.
- Lisi, F.A., and D. Malerba. 2002. "Spada: A spatial association discovery system." In *International Conference on Data Mining*. Bologna, Italy. 157-166.
- Liu, B. , W. Hsu, and Y. Ma. 1999. "Mining association rules with multiple minimum supports." In *Knowledge Discovery and Data Mining Conference*. 337-341.
- Lochner, L. 2004. Education, work, and crime: A human capital approach. *International Economic Review* 45.
- Lochner, L., and E. Moretti. 2004. The effect of education on crime: Evidence from prison inmates, arrests, and self-reports. *American Economic Review* 94.
- Lu, Yongmei, and Jean-Claude Thill. 2003. Assessing the cluster correspondence between paired point locations. *Geographical Analysis* 35 (4): 290-309.

- Lynch, Kevin. 1960. *The image of the city*. Cambridge, Mass.: MIT Press.
- MacEachren, A., ed. 1994. *Visualization in modern cartography: Setting the agenda*. Edited by D. R. F. Taylor and A.M. MacEachren, *Visualization in modern cartography*. Oxford, UK: Pergamon.
- MacEachren, A., and M-J. Kraak. 2001. Research challenges in geovisualization. *Cartography and Geographic Information Science*: 283–312.
- MacEachren, A. M., and M. J. Kraak. 1997. Exploratory cartographic visualization: Advancing the agenda. *Computers and Geosciences* 23: 335–343.
- MacEachren, A. M., M. Wachowicz, R. Edsall, D. Haug, and R. Masters. 1999. Constructing knowledge from multivariate spatiotemporal data: Integrating geographical visualization with knowledge discovery in database methods. *International Journal of Geographical Information Science* 13 (4): 311-334.
- Malczewski, J. , and A. Poetz. 2005. Residential burglaries and neighborhood socioeconomic context in london, ontario: Global and local regression analysis. *Professional Geographer* 57: 516–529.
- Malerba, D. 2008. A relational perspective on spatial data mining. *International Journal of Data Modelling and Management* 1 (1): 103-118.
- Malerba, D., F. Esposito, Lisi F.A., and A. Appice. 2002. Mining spatial association rules in census data. *Research in Official Statistics* (1): 19-45.
- Malerba, D., F. Esposito, A. Lanza, and F. Lisi. 2010. "Discovering geographic knowledge: The ingens system foundations of intelligent systems." In, eds. Z. Ras and S. Ohsuga: Springer Berlin / Heidelberg. 225-234.
- Malerba, D., A. Lanza, and A. Appice, eds. 2009. *Leveraging the power of spatial data mining to enhance the applicability of gis technology*. Edited by J. Han and R Cohen, *Geographic knowledge discovery and data mining, 2nd edition*: Taylor and Francis Group.
- Malerba, D., F.A. Lisi, A. Appice, and F. Sblendorio. 2003. "Mining census and geographic data in urban planning environments " In *Atti della Terza Conferenza Nazionale su Informatica e Pianificazione Urbana e Territoriale (INPUT 2003)*, eds. L. Santini and D. Zotta. Firenze, Italy.
- Malleson, N., A. Heppenstall, and L. See. 2009. An agent-based model of burglary. *Environment and Planning B: Planning and Design* 36 (3): 1103 -1123.
- Mannila, H., and H. Toivonen. 1997. Levelwise search and borders of theories in knowledge discovery. *Data Mining and Knowledge Discovery* 1 (3): 241-258.

- Manning, A., and J. Keane. 2001. Data allocation algorithm for parallel association rule discovery. *Lecture Notes in Computer Science* 2035.
- Matthews, S. A., T-C. Yang, K.L. Hayslett, and B. R. Ruback. 2010. Built environment and property crime in seattle, 1998 - 2000: A bayesian analysis. *Environment and Planning A* 42: 1403 - 1420.
- Maxfield, M. G. , and R. V. Clarke. 2004. "Understanding and preventing car theft." In *Crime Prevention Studies*: Monsey, NY: Criminal Justice Press.
- May, M., and A. Savinov. 2003. Spin!—an enterprise architecture for spatial data mining. *Lecture Notes in Computer Science* 2773 510-517.
- McCaghy, C. H., P. C. Giordano, and T. K. Henson. 1977. Auto theft: Offense and offender characteristics. *Criminology* 15 (3): 367-385.
- McGhey, R., ed. 1986. *Economic conditions, neighborhood organization, and urban crime*. Edited by A. Reiss and M. Tonry, *Communities and crimes*. Chicago: University of Chicago Press.
- Mennis, J., and D. Guo. 2009. Spatial data mining and geographic knowledge discovery - an introduction. *Computers, Environment and Urban Systems* 33 (6): 403-408.
- Mennis, J., and J.W. Liu. 2005. Mining association rules in spatio-temporal data: An analysis of urban socioeconomic and land cover change. *Transactions in GIS* 9 (1): 5-17.
- Miethe, T., and R. C. McCorkle. 2001. *Crime profiles: The anatomy of dangerous persons, places, and situations (2nd ed.)*. Los Angeles: Roxbury.
- Miller, H., ed. 2007. *Geographic data mining and knowledge discovery*. Edited by Wilson J.P. and A. S. Fotheringham, *Handbook of geographic information science*: Blackwell.
- Miller, H., and J. Han, eds. 2009. *Geographic data mining and knowledge discovery: An overview*. Edited by H. Miller and J. Han, *Geographic data mining and knowledge discovery*: CRC Press, Taylor and Francis Group.
- Miller, H. J., and E. A. Wentz. 2003. Representation and spatial analysis in geographic information systems. *Annals of the Association of American Geographers* 93: 574-594.
- Miller, H.J., and J. Han. 2001. *Geographical data mining and knowledge discover*. New York: Taylor & Francis Inc.
- Monmonier, M. 1989. Geographic brushing: Enhancing exploratory analysis of the scatterplot matrix. *Geographical Analysis* 21 (1): 81-84.
- Moran, P. A. P. . 1948. The interpretation of statistical maps. *Journal of the Royal Statistical Society. Series B (Methodological)* 10 (2): 243-251.

- Morenoff, J., and R.J. Sampson. 1997. Violent crime and the spatial dynamics of neighborhood transition: Chicago 1970–1990. *Social Forces* 76 (31–64).
- Morenoff, J.D., R.J. Sampson, and S.W. Raudenbush. 2001. Neighbourhood inequality, collective efficacy, and the spatial dynamics of urban violence. *Criminology*: 517–559.
- Mukerjee, A., and G Joe. 1990. "A qualitative model for space." In *AAAI-90 Workshop on Qualitative Vision*: AAAI Press. 721–727.
- Murray, A.T., I McGuffog, J.S. Western, and P. Mullins. 2001. Exploring spatial data analysis techniques for examining urban crime. *British Journal of Criminology* 41 (2): 309-329.
- Ng, R., ed. 2001. *Detecting outliers from large datasets*. Edited by H.J. Miller and J. Han, *Geographic data mining and knowledge discovery*. London: Taylor and Francis.
- Oatley, G., and B. Ewart. 2011. Data mining and crime analysis. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 1 (2): 147-153.
- Omiecinski, E. 2003. Alternative interest measures for mining associations in databases. *IEEE Transactions on Knowledge and Data Engineering* 15 (1): 57-69.
- Openshaw, S. 1983. *The modifiable areal unit problem*. Norwick: Geo Books.
- Openshaw, S., M. Charlton, C. Wymer, and A. Craft. 1987. A mark 1 geographical analysis machine for the automated analysis of point data sets. *International Journal of Geographical Information Systems* 1 (4): 335-358.
- Ord, J. K., and Arthur Getis. 1995. Local spatial autocorrelation statistics: Distributional issues and an application. *Geographical Analysis* 27 (4): 286-306.
- Ord, J. Keith, and Arthur Getis. 2001. Testing for local spatial autocorrelation in the presence of global autocorrelation. *Journal of Regional Science* 41 (3): 411-432.
- Ord, K. 1975. Estimation methods for models of spatial interaction. *Journal of the American Statistical Association* 70 (349): 120-126.
- Parthasarathy, S. 2002. "Efficient progressive sampling for association rules." Presented at the Proceedings of the 2002 IEEE International Conference on Data Mining.
- Parthasarathy, S. , M. J. Zaki, M. Ogihara, and W. Li. 2001. Parallel data mining for association rules on shared memory systems. *Knowledge and Information Systems* 3 (1): 1-29.
- Patterson, E.B. 1991. Poverty, inequality, and community crime rates. *Criminology* 29: 755-776.

- Phillips, P., and I. Lee. 2006. "Mining positive associations of urban criminal activities using hierarchical crime hot spots intelligence and security informatics." In, eds. H. Chen, F. Wang, C. Yang, D. Zeng, M. Chau and K. Chang: Springer Berlin / Heidelberg. 127-132.
- Phillips, P., and I. Lee. 2009. "Mining top-k and bottom-k correlative crime patterns through graph representations." In *Intelligence and Security Informatics, 2009. ISI '09. IEEE International Conference on*. 25-30.
- Phillips, P.; Lee, I. 2007. "Areal aggregated crime reasoning through density tracing." In *The 7th IEEE International Conference on Data Mining*
- Phillips, P.; Lee, I. 2009. "Criminal cross correlation mining and visualization_2009." In *Pacific Asia Workshop on Intelligence and security informatics (PAISI, 2009)*. Bangkok, Thailand.
- Phillips, P.; Lee, I. 2011. Crime analysis through spatial areal aggregated density patterns. *GeoInformatica* 15 (1): 49-74.
- Potchak, M., J.M. McGloin, and K.M. Zgoba. 2002. A spatial analysis of criminal effort: Auto theft in newark, new jersey. *Criminal Justice Policy Review* 13 (3): 257-285.
- Ratcliffe, J. H., and M. J. McCullagh. 1999. Hotbeds of crime and the search for spatial accuracy. *Journal of Geographical Systems* 1 (4): 385-398.
- Regoeczi, W. C. 2003. When context matters: A multilevel analysis of household and neighborhood crowding on aggression and withdrawal. *Journal of Environmental Psychology* 23 (4): 457-470.
- Rengert, G., ed. 1989. *Behavioral geography and criminal behavior*. Edited by D. J. Evans and D. T. Herbert, *The geography of crime*. London: Routledge.
- Rengert, G.F., and J. Wasilchick. 2000. *Suburban burglary: A tale of two suburbs*. Springfield, IL: C.C. Thomas Publishing.
- Rice, K. J., and W. R. Smith. 2002. Socioecological models of automotive theft: Integrating routine activity and social disorganization approaches. *Journal of Research in Crime and Delinquency* 39 (3): 304-336.
- Richard, M. 1986. Economic conditions, neighborhood organization, and urban crime. *Crime and Justice* 8 (Communities and Crime): 231-270.
- Ripley, B. D. 1977. Modelling spatial patterns. *Journal of the Royal Statistical Society Series B* 39: 172-192.

- Roberts, A., and S. Block. 2013. Explaining temporary and permanent motor vehicle theft rates in the united states a crime-specific approach *Journal of Research in Crime and Delinquency* 50 (3): 445-471.
- Roddick, J. F., and B. Lees, eds. 2001. *Paradigms for spatial and spatio-temporal data minig*. Edited by H. J. Miller and J. Han, *Geographic data mining and knowledge discovery*. London: Taylor and Francis.
- Rogerson, P. A., and P. Kedron. 2012. Optimal weights for focused tests of clustering using the local moran statistic. *Geographical Analysis* 44 (2): 121-133.
- Roosta, S. H. 2000. *Parallel processing and parallel algorithms : Theory and computation*. New York: Springer.
- Rossmo, D. K., ed. 1995. *Place, space, and police investigations: Hunting serial violent criminals*. Edited by J. E. Eck and D. A. Weisburd. Vol. 4, *Crime and place: Crime prevention studies*. Monsey, NY: Criminal Justice Press.
- Royaltey, H., E. Astrachan, and R. Sokal. 1975. Test for patterns in geographic variation. *Geographical Analysis* 7: 369-396.
- Sampson, R. J. 1985. Neighborhood and crime: The structural determinants of personal victimization. *Journal of Research in Crime and Delinquency* 22: 7-40.
- Sampson, R. J., J. Morenoff, and F. Earls. 1999. Beyond social capital: Neighborhood mechanism and structural sources of collective efficacy for children. *American Sociological Review* 64: 633-660.
- Sampson, R. J., and S. Raudenbush. 2004. The social structure of seeing disorder. *Social Psychology Quarterly* 67 (4): 319-342.
- Sampson, R. J., and S. Raudenbush, eds. 2001. *Disorder in urban neighborhoods - does it lead to crime?*, *National institute of justice research in brief*. Washinton D.C.: U.S. Department of Justice.
- Sampson, R.J., and W.B. Groves. 1989. Community structure and crime: Testing social disorganization theory. *American Journal of Sociology* 94 (4): 774-802.
- Sampson, R.J., S.W. Raudenbush, and F. Earls. 1997. Neighborhoods and violent crime: A multilevel study of collective efficacy. *Science* 277 (5328): 918-924.
- Sathyaraj, S.R.; Thangavelu, A.; Balasubramanian, S.; Sridhar, R., and M.; Prashanthi Devi Chandran, M. 2010. "Clustered spatial association rule to explore large volumes of georeferenced crime to crime data." In *The 3rd International conference on Cartography and GIS*. Nessebar, Bulgaria.

- Sen, A. 1976. Large sample-size distribution of statistics used in testing for spatial correlation. *Geographical Analysis* 8 (2): 175-184.
- Sen, A. , and S. Soot. 1977. Rank tests for spatial autocorrelation. *Environment and Planning A* 9: 897-903.
- Shaw, C. R., and H.D. McKay. 1942. *Juvenile delinquency and urban areas*. Chicago: University of Chicago Press.
- Shekhar, S., and S. Chawla. 2003. *Spatial databases: A tour*. Upper Saddle River, N. J: Prentice-Hall.
- Shekhar, S., C. T. Lu, X. Tan, S. Chawla, and R. R. Vatsavai, eds. 2001. *Map cube: A visualization tool for spatial data warehouses*. Edited by H. J. Miller and J. Han, *Geographic data mining and knowledge discovery*. London: Taylor and Francis.
- Shekhar, S., C.T. Lu, and P. Zhang. 2003a. A unified approach to detecting spatial outliers. *GeoInformatica* 7: 139-166.
- Shekhar, S., P.R. Schrater, R.R. Vatsavai, W. Wu, and S. Chawla. 2002. Spatial contextual classification and prediction models for mining geospatial data. *IEEE Transactions on Multimedia* 4 (2): 174 - 188
- Shekhar, S., P. Zhang, Y. Huang, and R.R. Vatsavai, eds. 2003b. *Trends in spatial data mining*. Edited by H.; Joshi Kargupta, A. , *Data mining: Next generation challenges and future directions*: AAAI/MIT Press.
- Simoff, S., M. Böhlen, and A. Mazeika. 2008a. *Visual data mining: Theory, techniques and tools for visual analytics*. Springer Berlin / Heidelberg.
- Simoff, S.J., M.H. Böhlen, and A. Mazeika. 2008b. "Visual data mining: An introduction and overview." In *Visual data mining: Theory, techniques and tools for visual analytics*, eds. S.J. Simoff, M.H. Böhlen and A. Mazeika: Springer Berlin / Heidelberg. 1-12.
- Simoff, S.J., M.H. Böhlen, and A. Mazeika, eds. 2008c. *Preface*. Edited by S.J. Simoff, M.H. Böhlen and A. Mazeika. Vol. 4404, *Visual data mining: Theory, techniques and tools for visual analytics, lecture notes in computer science*.
- Srikant, R., and R. Agrawal. 1995. "Mining generalized association rules." Presented at the Proceedings of the 21th International Conference on Very Large Data Bases.
- Srikant, R., and R. Agrawal. 1996. Mining quantitative association rules in large relational tables. *SIGMOD Rec.* 25 (2): 1-12.
- Stakhovych, S., and T. H. A. Bijmolt. 2009. Specification of spatial models: A simulation study on weights matrices. *Papers in Regional Science* 88 (2): 389-408.

- Statistica. 2012. "Association rules - graphical representation of associations." <http://documentation.statsoft.com/STATISTICAHelp.aspx?path=GXX/AssociationRules/Overviews/AssociationRulesGraphicalRepresentationofAssociations> 2012).
- Stetzer, F. 1982. Specifying weights in spatial forecasting models: The results of some experiments. *Environment and Planning A* 14 (5): 571-584.
- Student. 1914. The elimination of spurious correlation due to position in time or space. *Biometrika* 10 (1): 179-180.
- Suresh, G., and R. Tewksbury. 2013. Locations of motor vehicle theft and recovery. *American Journal of Criminal Justice* 38 (2): 200-215.
- Tan, P-N., M. Steinbach, and V. Kumar. 2006. *Introduction to data mining*. Pearson Addison Wesley.
- Thomas, J., and K. Cook. 2005. *Illuminating the path: Research and development agenda for visual analytics*. Los Alamitos: IEEE Press
- Toivonen, H. 1996. "Sampling large databases for association rules." Presented at the Proceedings of the 22th International Conference on Very Large Data Bases.
- Tolnay, S.E., G. Deane, and E.M. Beck. 1996. Vicarious violence: Spatial effects on southern lynchings, 1890–1919. *American Journal of Sociology* 102: 788–815.
- Townsley, M. 2009. Spatial autocorrelation and impacts on criminology. *Geographical Analysis* 41 (4): 452-461.
- Tukey, J. 1977. *Exploratory data analysis*. Addison-Wesley.
- Verma, K., O. Vyas, and R. Vyas. 2005. "Temporal approach to association rule mining using t-tree and p-tree machine learning and data mining in pattern recognition." In, eds. P. Perner and A. Imiya: Springer Berlin / Heidelberg. 639-639.
- Walsh, J. A., and R. B. Taylor. 2007a. Predicting decade-long changes in community motor vehicle theft rates: Impacts of structure and surround. *Journal of Research in Crime and Delinquency* 44 (1): 64-90.
- Walsh, J., and R. B. Taylor. 2007b. Community structural predictors of spatially aggregated motor vehicle theft rates: Do they replicate? *Journal of Criminal Justice* 35: 297-311.
- Wang, C., and C. Tjortjis. 2004. Prices: An efficient algorithm for mining association rules. *Lecture Notes in Computer Science* 3177: 352 - 358.
- Wang, Dawei, Wei Ding, Henry Lo, Melissa Morabito, Ping Chen, Josue Salazar, and Tomasz Stepinski. 2013. Understanding the spatial distribution of crime based on its

- related variables using geospatial discriminative patterns. *Computers, Environment and Urban Systems* 39 (0): 93-106.
- Wang, Y., K. Kockelman, and X. Wang. 2012. "The impact of weight matrices on parameter estimation and inference: A case study of binary response using land use data." In *The 91st annual meeting of the Transportation Research Board*. Washington, DC.
- Warner, B. D., and G. Pierce. 1993. Reexamining social disorganization theory using calls to the police as a measure of crime. *Criminology* 31 (3): 493-517.
- Wartenberg, D. 1985. Multivariate spatial correlation: A method for exploratory geographical analysis. *Geographical Analysis* 17 (4): 263-283.
- Widener, M. J., N. C. Crago, and J. Aldstadt. 2012. Developing a parallel computational implementation of amoeba. *International Journal of Geographical Information Science*: 1-17.
- Wiles, P., and A. Costello. 2000. *The 'road to nowhere': The evidence for travelling criminals*. London: Research, Development and Statistics Directorate
- Williamson, D., S. McLafferty, V. Goldsmith, J. Mollenkop, and P. McGuire. 1999. A better method to smooth crime incident data. *ESRI ArcUser Magazine* January-March 1999: 1-5.
- Wilson, W. 1987. *The truly disadvantaged*. Chicago: University of Chicago Press.
- Wojciechowski, M., and M. Zakrzewicz. 2002. "Dataset filtering techniques in constraint-based frequent pattern mining pattern detection and discovery." In, eds. D. Hand, N. Adams and R. Bolton: Springer Berlin / Heidelberg. 301-318.
- Wong, P-C., P. Whitney, and J. Thomas. 1999. "Visualizing association rules for text mining." In *Information Visualization, 1999. (Info Vis '99) Proceedings. 1999 IEEE Symposium on*. 120-123, 152.
- Wright, R., and S. Decker. 1994. *Burglary on the job*. Boston: Northeastern University Press.
- Yan, J., and J. C. Thill. 2009. Visual data mining in spatial interaction analysis with self-organizing maps. *Environment and Planning B: Planning and Design* 36 (3): 466-486.
- Yang, L. 2008. "Visual exploration of frequent itemsets and association rules." In *Visual data mining: An introduction and overview*, eds. S. J. Simoff, M. H. Bohlen and A. Mazeika: Springer-Verlag. 60-75.
- Young, F. W. 1987. *Multidimensional scaling: History, theory, and applications*. Lawrence Erlbaum Associates.

- Yuan, Y., and T. Huang. 2005. A matrix algorithm for mining association rules. *Lecture Notes in Computer Science* 3644: 370-379.
- Zaki, M.J., S. Parthasarathy, M. Ogihara, and W. Li. 1997. Parallel algorithms for fast discovery of association rules. *Data Mining and Knowledge Discovery (special issue on scalable high-performance computing for KDD)* 1 (4): 343-373.
- Zhang, H., G. Suresh, and Y. Qiu. 2012. Issues in the aggregation and spatial analysis of neighborhood crime. *Annals of GIS* 18 (3): 173-183.
- Zhang, X., and M. Pazner. 2004. The icon imagemap technique for multivariate geospatial data visualization: Approach and software system. *Cartography and Geographic Information Science* 31 (1): 29-41.

APPENDIX A: REPRESENTATIVE MINED SARS FOR DANGEROUS STREETS
DUE TO CRIME OF ALL TYPES USING CRISP MAPPING, SUPPORT
THRESHOLD OF 5%, CONFIDENCE THRESHOLD OF 40%

	Rules	C%	S%
1	HeterogeneityInx=H UnstableRent=H HEdu=M Income=L ==> CAT=H SingleParentF=H HomeOwn=L Male17-24=H HeterogeneityInx=H ==>	5	81
2	CAT=H	5	81
3	SingleParentF=H HeterogeneityInx=H MUHouse=H Income=L ==> CAT=H	5	80
4	SingleParentF=H UnstableRent=H MUHouse=H Income=L ==> CAT=H	5	79
5	SingleParentF=H HeterogeneityInx=H MUHouse=H ==> CAT=H	5	79
6	SingleParentF=H Male17-24=H MUHouse=H ==> CAT=H	5	78
7	HeterogeneityInx=H UnstableRent=H MUHouse=H Income=L ==> CAT=H	6	77
8	SingleParentF=H UnstableRent=H MUHouse=H ==> CAT=H	6	77
9	SingleParentF=H MUHouse=H ==> CAT=H	7	76
10	SingleParentF=H MUHouse=H Income=L ==> CAT=H SingleParentF=H HeterogeneityInx=H UnstableRent=H Income=L ==>	6	76
11	CAT=H	5	76
12	SingleParentF=H HomeOwn=L MUHouse=H ==> CAT=H	5	76
13	AfricanA=H MUHouse=H Income=L ==> CAT=H	5	75
14	HomeOwn=L Male17-24=H HeterogeneityInx=H Income=L ==> CAT=H	6	75
15	HomeOwn=MH UnstableRent=M MUHouse=M ==> CAT=MH	6	75
16	UnstableRent=H HEdu=M MUHouse=H Income=L ==> CAT=H	5	74
17	AfricanA=H MUHouse=H ==> CAT=H	5	74
18	Male17-24=H HeterogeneityInx=H UnstableRent=H Income=L ==> CAT=H	5	74
19	SingleParentF=H HomeOwn=L HeterogeneityInx=H Income=L ==> CAT=H	6	74
20	HomeOwn=L Male17-24=H HEdu=M Income=L ==> CAT=H	6	73
21	SingleParentF=H HomeOwn=L HeterogeneityInx=H ==> CAT=H	6	73
22	SingleParentF=H HomeOwn=L Male17-24=H HEdu=M ==> CAT=H	5	73
23	SingleParentF=H HeterogeneityInx=H UnstableRent=H ==> CAT=H	5	72
24	SingleParentF=H HomeOwn=L HEdu=M Income=L ==> CAT=H	7	72
25	SingleParentF=H HomeOwn=L AfricanA=H HEdu=M ==> CAT=H	6	72
26	HeterogeneityInx=H UnstableRent=H HEdu=M ==> CAT=H	6	71
27	Male17-24=H HeterogeneityInx=H MUHouse=H Income=L ==> CAT=H	5	71
28	HomeOwn=L HeterogeneityInx=H HEdu=M Income=L ==> CAT=H	5	71
29	HomeOwn=L HeterogeneityInx=H AfricanA=H ==> CAT=H	5	71
30	SingleParentF=H HomeOwn=L HEdu=M ==> CAT=H	7	71
31	HeterogeneityInx=H UnstableRent=H Income=L ==> CAT=H	7	70
32	HomeOwn=L HeterogeneityInx=H UnstableRent=H Income=L ==> CAT=H	5	70
33	SingleParentF=H HomeOwn=L UnstableRent=H HEdu=M ==> CAT=H	5	70
34	SingleParentF=H HomeOwn=L Male17-24=H Income=L ==> CAT=H	7	70

APPENDIX A: (continued)

35	Male17-24=H UnstableRent=H MUHouse=H Income=L ==> CAT=H	6	69
36	SingleParentF=H HomeOwn=L Male17-24=H ==> CAT=H	7	69
37	SingleParentF=H HomeOwn=L Male17-24=H AfricanA=H ==> CAT=H	5	68
38	HomeOwn=MH HeterogeneityInx=MH ==> CAT=MH	5	68
39	HomeOwn=MH Male17-24=H ==> CAT=MH	6	68
40	HeterogeneityInx=H MUHouse=H Income=L ==> CAT=H	6	67
41	HomeOwn=L Male17-24=H HeterogeneityInx=H ==> CAT=H	6	67
42	HomeOwn=MH UnstableRent=M ==> CAT=MH	8	67
43	Male17-24=H UnstableRent=H HEdu=M Income=L ==> CAT=H	6	66
44	HomeOwn=L Male17-24=H UnstableRent=H HEdu=M ==> CAT=H	5	66
45	SingleParentF=H HomeOwn=L Male17-24=H UnstableRent=H ==> CAT=H	5	66
46	Male17-24=H UnstableRent=M ==> CAT=MH	7	66
47	UnstableRent=H MUHouse=H Income=L ==> CAT=H	8	65
48	HeterogeneityInx=H UnstableRent=H MUHouse=H ==> CAT=H	6	65
49	HomeOwn=L AfricanA=H HEdu=M ==> CAT=H	6	65
50	HomeOwn=L AfricanA=H HEdu=M Income=L ==> CAT=H	6	65
51	HomeOwn=L HeterogeneityInx=H Income=L ==> CAT=H	7	65
52	HomeOwn=L Male17-24=H MUHouse=H Income=L ==> CAT=H	5	65
53	HomeOwn=L Male17-24=H HEdu=M ==> CAT=H	6	65
54	HomeOwn=L Male17-24=H AfricanA=H ==> CAT=H	6	65
55	HomeOwn=L Male17-24=H HeterogeneityInx=H UnstableRent=H ==> CAT=H	5	65
56	SingleParentF=H HomeOwn=L ==> CAT=H	10	65
57	SingleParentF=H HomeOwn=L Income=L ==> CAT=H	9	65
58	SingleParentF=H HomeOwn=L AfricanA=H UnstableRent=H ==> CAT=H	5	65
59	HomeOwn=MH MUHouse=M ==> CAT=MH	7	65
60	HomeOwn=L HeterogeneityInx=H UnstableRent=H MUHouse=H ==> CAT=H	5	64
61	HomeOwn=L Male17-24=H AfricanA=H Income=L ==> CAT=H	6	64
62	SingleParentF=H HomeOwn=L UnstableRent=H Income=L ==> CAT=H	6	64
63	SingleParentF=H HomeOwn=L AfricanA=H ==> CAT=H	8	64
64	SingleParentF=H HomeOwn=L AfricanA=H Income=L ==> CAT=H	7	64
65	UnstableRent=M MUHouse=M Income=L ==> CAT=MH	5	64
66	AfricanA=H UnstableRent=M ==> CAT=MH	6	64
67	HEdu=M MUHouse=H Income=L ==> CAT=H	6	63
68	UnstableRent=H HEdu=M MUHouse=H ==> CAT=H	6	63
69	Male17-24=H MUHouse=H Income=L ==> CAT=H	6	63
70	Male17-24=H HeterogeneityInx=H UnstableRent=H MUHouse=H ==> CAT=H	5	63
71	HomeOwn=L UnstableRent=H MUHouse=H Income=L ==> CAT=H	6	63
72	HomeOwn=L UnstableRent=H HEdu=M Income=L ==> CAT=H	6	63
73	HomeOwn=L HeterogeneityInx=H UnstableRent=H ==> CAT=H	6	63

APPENDIX A: (continued)

74	SingleParentF=H UnstableRent=H HEdu=M ==> CAT=H	6	63
75	SingleParentF=H UnstableRent=H HEdu=M Income=L ==> CAT=H	6	63
76	SingleParentF=H AfricanA=H UnstableRent=H HEdu=M ==> CAT=H	5	63
77	SingleParentF=H HomeOwn=L UnstableRent=H ==> CAT=H	7	63
78	AfricanA=H UnstableRent=M Income=L ==> CAT=MH	5	63
79	HomeOwn=MH HEdu=L ==> CAT=MH	6	63
80	HomeOwn=MH AfricanA=L UnstableRent=M ==> CAT=MH	5	63
81	HomeOwn=L HeterogeneityInx=H HEdu=M ==> CAT=H	5	62
82	HomeOwn=L Male17-24=H Income=L ==> CAT=H	8	62
83	Employment=M UnstableRent=M ==> CAT=MH	8	62
84	HomeOwn=MH Employment=M ==> CAT=MH	7	62
85	HeterogeneityInx=H UnstableRent=H ==> CAT=H	7	61
86	Male17-24=H HeterogeneityInx=H MUHouse=H ==> CAT=H	6	61
87	Male17-24=H HeterogeneityInx=H UnstableRent=H ==> CAT=H	6	61
88	HomeOwn=L HeterogeneityInx=H MUHouse=H ==> CAT=H	6	61
89	SingleParentF=H Male17-24=H UnstableRent=H Income=L ==> CAT=H	6	61
90	HeterogeneityInx=H UnstableRent=M Income=L ==> CAT=MH	6	61
91	AfricanA=H UnstableRent=H HEdu=M Income=L ==> CAT=H	6	60
92	Male17-24=H UnstableRent=H HEdu=M ==> CAT=H	6	60
93	HomeOwn=L UnstableRent=H HEdu=M MUHouse=H ==> CAT=H	5	60
94	HomeOwn=L AfricanA=H ==> CAT=H	9	60
95	HomeOwn=L AfricanA=H UnstableRent=H ==> CAT=H	6	60
96	SingleParentF=H UnstableRent=H Income=L ==> CAT=H	8	60
97	SingleParentF=H Male17-24=H UnstableRent=H ==> CAT=H	6	60
98	Male17-24=H MUHouse=M ==> CAT=MH	6	60
99	Male17-24=H MUHouse=M Income=L ==> CAT=MH	5	60
100	Male17-24=H UnstableRent=M Income=L ==> CAT=MH	5	60
101	SingleParentF=H HomeOwn=MH Income=L ==> CAT=MH	5	60
102	SingleParentF=L HomeOwn=MH AfricanA=L ==> CAT=MH	6	60
103	MUHouse=H Income=L ==> CAT=H	9	59
104	UnstableRent=H HEdu=M Income=L ==> CAT=H	8	59
105	AfricanA=H UnstableRent=H HEdu=M ==> CAT=H	6	59
106	HeterogeneityInx=H MUHouse=H ==> CAT=H	7	59
107	Male17-24=H AfricanA=H UnstableRent=H Income=L ==> CAT=H	5	59
108	HomeOwn=L MUHouse=H Income=L ==> CAT=H	7	59
109	HomeOwn=L HEdu=M Income=L ==> CAT=H	9	59
110	HomeOwn=L AfricanA=H Income=L ==> CAT=H	8	59
111	HomeOwn=L AfricanA=H UnstableRent=H Income=L ==> CAT=H	5	59
112	HomeOwn=L HeterogeneityInx=H ==> CAT=H	8	59
113	HomeOwn=L Male17-24=H UnstableRent=H Income=L ==> CAT=H	6	59
114	SingleParentF=H AfricanA=H UnstableRent=H Income=L ==> CAT=H	6	59

APPENDIX A: (continued)

115	HeterogeneityInx=H UnstableRent=M ==> CAT=MH	8	59
116	HeterogeneityInx=MH MUHouse=M ==> CAT=MH	5	59
117	SingleParentF=H HomeOwn=MH ==> CAT=MH	5	59
118	SingleParentF=L HomeOwn=MH ==> CAT=MH	6	59
119	Male17-24=H UnstableRent=H Income=L ==> CAT=H	7	58
120	SingleParentF=H UnstableRent=H ==> CAT=H	8	58
121	SingleParentF=H AfricanA=H UnstableRent=H ==> CAT=H	6	58
122	AfricanA=H MUHouse=M Income=L ==> CAT=MH	6	58
123	HomeOwn=MH Income=L ==> CAT=MH	8	58
124	HomeOwn=MH AfricanA=L ==> CAT=MH	7	58
125	SingleParentF=H MUHouse=M Income=L ==> CAT=MH	5	58
126	HomeOwn=L Employment=M Income=L ==> CAT=H	5	57
127	MUHouse=M Income=L ==> CAT=MH	9	57
128	AfricanA=H MUHouse=M ==> CAT=MH	6	57
129	HeterogeneityInx=MH Employment=M ==> CAT=MH	5	57
130	HomeOwn=MH ==> CAT=MH	13	57
131	SingleParentF=H UnstableRent=M ==> CAT=MH	5	57
132	AfricanA=H UnstableRent=H Income=L ==> CAT=H	7	56
133	Male17-24=H UnstableRent=H MUHouse=H ==> CAT=H	6	56
134	Male17-24=H AfricanA=H UnstableRent=H ==> CAT=H	5	56
135	HomeOwn=L UnstableRent=H HEdu=M ==> CAT=H	7	56
136	HomeOwn=L Male17-24=H UnstableRent=H MUHouse=H ==> CAT=H	5	56
137	HeterogeneityInx=H MUHouse=M ==> CAT=MH	5	56
138	SingleParentF=H MUHouse=M ==> CAT=MH	6	56
139	SingleParentF=H UnstableRent=M Income=L ==> CAT=MH	5	56
140	SingleParentF=H AfricanA=H MUHouse=L Income=L ==> CAT=MH	5	56
141	AfricanA=H UnstableRent=H ==> CAT=H	7	55
142	HomeOwn=L HEdu=M MUHouse=H ==> CAT=H	5	55
143	HomeOwn=L UnstableRent=H Income=L ==> CAT=H	9	55
144	HomeOwn=L Male17-24=H MUHouse=H ==> CAT=H	5	55
145	UnstableRent=M MUHouse=M ==> CAT=MH	10	55
146	Employment=M MUHouse=H ==> CAT=H	5	54
147	HomeOwn=L Male17-24=H ==> CAT=H	9	54
148	UnstableRent=M Income=M ==> CAT=MH	5	54
149	UnstableRent=M Income=L ==> CAT=MH	9	54
150	SingleParentF=H MUHouse=L Income=L ==> CAT=MH	7	54
151	SingleParentF=H AfricanA=H MUHouse=L ==> CAT=MH	5	54
152	UnstableRent=H Income=L ==> CAT=H	11	53
153	UnstableRent=H HEdu=M ==> CAT=H	8	53
154	HomeOwn=L Income=L ==> CAT=H	13	53
155	HomeOwn=L Male17-24=H UnstableRent=H ==> CAT=H	6	53

APPENDIX A: (continued)

156	UnstableRent=M ==> CAT=MH	17	53
157	UnstableRent=M HEdu=L ==> CAT=MH	6	53
158	HomeOwn=MH HEdu=M ==> CAT=MH	7	53
159	SingleParentF=H MUHouse=L ==> CAT=MH	7	53
160	HEdu=M MUHouse=H ==> CAT=H	7	52
161	HomeOwn=L UnstableRent=H MUHouse=H ==> CAT=H	8	52
162	HeterogeneityInx=MH Income=L ==> CAT=MH	7	52
163	HomeOwn=MH Employment=H ==> CAT=MH	5	52
164	SingleParentF=H AfricanA=H Employment=M ==> CAT=MH	5	52
165	SingleParentF=H AfricanA=H Employment=M Income=L ==> CAT=MH	5	52
166	UnstableRent=H MUHouse=H ==> CAT=H	10	51
167	Employment=M UnstableRent=H ==> CAT=H	5	51
168	Male17-24=H MUHouse=H ==> CAT=H	7	51
169	SingleParentF=H Male17-24=H HeterogeneityInx=H Income=L ==> CAT=H	6	51
170	HEdu=L MUHouse=M ==> CAT=MH	5	51
171	AfricanA=MH ==> CAT=MH	7	51
172	HomeOwn=H HeterogeneityInx=H Income=L ==> CAT=MH	5	51
173	UnstableRent=M HEdu=M ==> CAT=MH	9	50
174	UnstableRent=M HEdu=M Income=L ==> CAT=MH	5	50
175	SingleParentF=H HeterogeneityInx=H AfricanA=H Income=L ==> CAT=MH	6	50
176	Male17-24=H UnstableRent=H ==> CAT=H	7	49
177	HomeOwn=L MUHouse=H ==> CAT=H	9	49
178	HomeOwn=L UnstableRent=H ==> CAT=H	11	49
179	Employment=H UnstableRent=M ==> CAT=MH	7	49
180	AfricanA=H MUHouse=L Income=L ==> CAT=MH	5	49
181	HeterogeneityInx=MH ==> CAT=MH	11	49
182	HomeOwn=MH Male17-24=L ==> CAT=MH	6	49
183	SingleParentF=L AfricanA=L UnstableRent=M ==> CAT=MH	7	49
184	SingleParentF=H Male17-24=H HeterogeneityInx=H ==> CAT=H	6	48
185	HEdu=M MUHouse=M Income=L ==> CAT=MH	5	48
186	HEdu=L Income=L ==> CAT=MH	7	48
187	Employment=M MUHouse=M ==> CAT=MH	6	48
188	Employment=L Income=L ==> CAT=MH	6	48
189	AfricanA=L UnstableRent=M ==> CAT=MH	9	48
190	Male17-24=H Employment=H ==> CAT=MH	7	48
191	SingleParentF=H AfricanA=H Income=L ==> CAT=MH	11	48
192	SingleParentF=H HeterogeneityInx=H AfricanA=H ==> CAT=MH	7	48
193	SingleParentF=L UnstableRent=M ==> CAT=MH	7	48
194	HomeOwn=L HEdu=M ==> CAT=H	9	47
195	MUHouse=M ==> CAT=MH	15	47
196	Employment=H MUHouse=M ==> CAT=MH	6	47

APPENDIX A: (continued)

197	Employment=M Income=L ==> CAT=MH	10	47
198	AfricanA=H Employment=M ==> CAT=MH	6	47
199	AfricanA=H Employment=M Income=L ==> CAT=MH	6	47
200	AfricanA=L UnstableRent=M MUHouse=M ==> CAT=MH	6	47
201	HeterogeneityInx=H MUHouse=L Income=L ==> CAT=MH	5	47
202	HeterogeneityInx=H HEdu=M MUHouse=L ==> CAT=MH	5	47
203	HeterogeneityInx=H AfricanA=H Income=L ==> CAT=MH	7	47
204	HeterogeneityInx=MH AfricanA=L ==> CAT=MH	6	47
205	HomeOwn=L HEdu=L ==> CAT=MH	6	47
206	SingleParentF=H Male17-24=H Income=L ==> CAT=H	8	46
207	SingleParentF=H Male17-24=H AfricanA=H Income=L ==> CAT=H	6	46
208	MUHouse=L Income=L ==> CAT=MH	9	46
209	HEdu=M MUHouse=L Income=L ==> CAT=MH	5	46
210	AfricanA=H Income=L ==> CAT=MH	13	46
211	AfricanA=L Employment=H UnstableRent=M ==> CAT=MH	5	46
212	HomeOwn=H UnstableRent=M ==> CAT=MH	5	46
213	SingleParentF=H Income=L ==> CAT=MH	14	46
214	SingleParentF=H Employment=M ==> CAT=MH	7	46
215	SingleParentF=H Employment=M Income=L ==> CAT=MH	7	46
216	SingleParentF=H AfricanA=H ==> CAT=MH	11	46
217	SingleParentF=H HeterogeneityInx=H Income=L ==> CAT=MH	8	46
218	SingleParentF=MH ==> CAT=MH	8	46
219	SingleParentF=L MUHouse=H ==> CAT=MH	5	46
220	SingleParentF=L UnstableRent=M MUHouse=M ==> CAT=MH	5	46
221	UnstableRent=H ==> CAT=H	12	45
222	Male17-24=H HeterogeneityInx=H HEdu=M ==> CAT=H	5	45
223	SingleParentF=H Male17-24=H ==> CAT=H	9	45
224	SingleParentF=H Male17-24=H HEdu=M Income=L ==> CAT=H	5	45
225	SingleParentF=H Male17-24=H AfricanA=H ==> CAT=H	7	45
226	Income=L ==> CAT=MH	23	45
227	Employment=H MUHouse=H ==> CAT=MH	6	45
228	AfricanA=H MUHouse=L ==> CAT=MH	6	45
229	HeterogeneityInx=L UnstableRent=M ==> CAT=MH	5	45
230	Male17-24=H ==> CAT=MH	15	45
231	Male17-24=H AfricanA=H ==> CAT=MH	8	45
232	Male17-24=H AfricanA=H Income=L ==> CAT=MH	7	45
233	HomeOwn=H Income=L ==> CAT=MH	6	45
234	HomeOwn=H HeterogeneityInx=H HEdu=M ==> CAT=MH	5	45
235	SingleParentF=H Male17-24=H AfricanA=H Income=L ==> CAT=MH	6	45
236	HomeOwn=L Employment=M ==> CAT=H	6	44
237	SingleParentF=H AfricanA=H HEdu=M Income=L ==> CAT=H	7	44

APPENDIX A: (continued)

238	SingleParentF=H Male17-24=H HEdu=M ==> CAT=H	5	44
239	Employment=M HEdu=L ==> CAT=MH	6	44
240	AfricanA=H ==> CAT=MH	14	44
241	AfricanA=L MUHouse=H ==> CAT=MH	6	44
242	HeterogeneityInx=H Income=L ==> CAT=MH	12	44
243	HeterogeneityInx=H Employment=M Income=L ==> CAT=MH	6	44
244	HeterogeneityInx=MH HEdu=M ==> CAT=MH	6	44
245	Male17-24=H Income=L ==> CAT=MH	12	44
246	Male17-24=H AfricanA=H HEdu=M ==> CAT=MH	5	44
247	Male17-24=L Income=L ==> CAT=MH	8	44
248	HomeOwn=H HeterogeneityInx=H ==> CAT=MH	8	44
249	SingleParentF=H ==> CAT=MH	14	44
250	SingleParentF=H HeterogeneityInx=H ==> CAT=MH	8	44
251	SingleParentF=H Male17-24=H Income=L ==> CAT=MH	8	44
252	MUHouse=H ==> CAT=H	11	43
253	Male17-24=H HeterogeneityInx=H Income=L ==> CAT=H	7	43
254	HeterogeneityInx=H MUHouse=L ==> CAT=MH	7	43
255	HeterogeneityInx=H AfricanA=H ==> CAT=MH	8	43
256	Male17-24=H HEdu=M Income=L ==> CAT=MH	7	43
257	Male17-24=L UnstableRent=M ==> CAT=MH	8	43
258	SingleParentF=H Male17-24=H AfricanA=H ==> CAT=MH	6	43
259	Employment=L Income=L ==> CAT=H	5	42
260	Male17-24=H Employment=M ==> CAT=H	5	42
261	Male17-24=H Employment=M Income=L ==> CAT=H	5	42
262	Male17-24=H AfricanA=H Income=L ==> CAT=H	7	42
263	Male17-24=H AfricanA=H HEdu=M ==> CAT=H	5	42
264	HomeOwn=L ==> CAT=H	15	42
265	SingleParentF=H HEdu=M Income=L ==> CAT=H	8	42
266	SingleParentF=H AfricanA=H HEdu=M ==> CAT=H	7	42
267	HEdu=M MUHouse=M ==> CAT=MH	8	42
268	UnstableRent=L Income=L ==> CAT=MH	6	42
269	UnstableRent=L MUHouse=L Income=L ==> CAT=MH	5	42
270	Employment=M HEdu=M Income=L ==> CAT=MH	6	42
271	Employment=L ==> CAT=MH	6	42
272	Male17-24=H Employment=M Income=L ==> CAT=MH	5	42
273	HomeOwn=H HeterogeneityInx=H MUHouse=L ==> CAT=MH	6	42
274	HomeOwn=L Employment=H ==> CAT=MH	6	42
275	HomeOwn=L HeterogeneityInx=L ==> CAT=MH	6	42
276	SingleParentF=H HEdu=M Income=L ==> CAT=MH	8	42
277	SingleParentF=H Male17-24=H ==> CAT=MH	8	42
278	SingleParentF=H Male17-24=H HEdu=M ==> CAT=MH	5	42

APPENDIX A: (continued)

279	SingleParentF=MH Income=L ==> CAT=MH	5	42
280	SingleParentF=L HomeOwn=L ==> CAT=MH	6	42
281	Male17-24=H AfricanA=H ==> CAT=H	7	41
282	SingleParentF=H HEdu=M ==> CAT=H	8	41
283	SingleParentF=H HeterogeneityInx=H Income=L ==> CAT=H	7	41
284	HEdu=M Income=L ==> CAT=MH	13	41
285	AfricanA=H HEdu=M Income=L ==> CAT=MH	8	41
286	HeterogeneityInx=H Employment=M ==> CAT=MH	7	41
287	Male17-24=H Employment=M ==> CAT=MH	5	41
288	HomeOwn=L Male17-24=L ==> CAT=MH	7	41
289	SingleParentF=H HEdu=M ==> CAT=MH	8	41
290	SingleParentF=H AfricanA=H HEdu=M ==> CAT=MH	7	41
291	SingleParentF=H AfricanA=H HEdu=M Income=L ==> CAT=MH	7	41
292	AfricanA=H HEdu=M Income=L ==> CAT=H	8	40
293	Male17-24=H HEdu=M Income=L ==> CAT=H	7	40
294	SingleParentF=H ==> CAT=H	12	40
295	SingleParentF=H Income=L ==> CAT=H	12	40
296	SingleParentF=H AfricanA=H Income=L ==> CAT=H	9	40
297	SingleParentF=H HeterogeneityInx=H ==> CAT=H	8	40
298	SingleParentF=H HeterogeneityInx=H HEdu=M ==> CAT=H	5	40
299	SingleParentF=H HeterogeneityInx=H AfricanA=H Income=L ==> CAT=H	5	40
300	HEdu=L ==> CAT=MH	14	40
301	UnstableRent=H ==> CAT=MH	11	40
302	Employment=H HEdu=L ==> CAT=MH	7	40
303	AfricanA=H HEdu=M ==> CAT=MH	9	40
304	HeterogeneityInx=H ==> CAT=MH	15	40
305	Male17-24=H HEdu=M ==> CAT=MH	8	40
306	Male17-24=H HeterogeneityInx=H Income=L ==> CAT=MH	7	40
307	SingleParentF=H HeterogeneityInx=H HEdu=M ==> CAT=MH	5	40

APPENDIX B: REPRESENTATIVE MINED SARS FOR DANGEROUS STREET
DUE TO MOTOR VEHICLE THEFT USING CRISP MAPPING, SUPPORT
THRESHOLD OF 5%, CONFIDENCE THRESHOLD OF 40%

	Rules	S%	C%
1	SingleParentF=H HeterogeneityInx=H MUHouse=H ==> MVT=H	5	73
2	HeterogeneityInx=H UnstableRent=H MUHouse=H Income=L ==> MVT=H	5	71
3	Male17-24=H HeterogeneityInx=H UnstableRent=H Income=L ==> MVT=H	5	71
4	HomeOwn=L Male17-24=H HeterogeneityInx=H MUHouse=H ==> MVT=H	5	71
5	HomeOwn=L Male17-24=H HeterogeneityInx=H ==> MVT=H HomeOwn=L Male17-24=H HeterogeneityInx=H UnstableRent=H ==>	6	70
6	MVT=H	5	70
7	Male17-24=H HeterogeneityInx=H MUHouse=H Income=L ==> MVT=H HomeOwn=L HeterogeneityInx=H UnstableRent=H MUHouse=H ==>	5	69
8	MVT=H	6	69
9	HomeOwn=L Male17-24=H HeterogeneityInx=H Income=L ==> MVT=H	5	69
10	HeterogeneityInx=H UnstableRent=H HEdu=M ==> MVT=H	5	68
11	HeterogeneityInx=H UnstableRent=H MUHouse=H ==> MVT=H	7	67
12	Male17-24=H HEdu=M MUHouse=H ==> MVT=H	5	67
13	HeterogeneityInx=H UnstableRent=H Income=L ==> MVT=H	6	66
14	Male17-24=H HeterogeneityInx=H MUHouse=H ==> MVT=H	6	66
15	Male17-24=H HeterogeneityInx=H UnstableRent=H ==> MVT=H Male17-24=H HeterogeneityInx=H UnstableRent=H MUHouse=H ==>	6	66
16	MVT=H	5	66
17	HomeOwn=L HeterogeneityInx=H MUHouse=H ==> MVT=H	6	66
18	HeterogeneityInx=H MUHouse=H Income=L ==> MVT=H	6	65
19	HomeOwn=L HeterogeneityInx=H UnstableRent=H ==> MVT=H	6	65
20	HomeOwn=L HeterogeneityInx=H UnstableRent=H Income=L ==> MVT=H	5	65
21	HeterogeneityInx=H MUHouse=H ==> MVT=H	7	63
22	HeterogeneityInx=H UnstableRent=H ==> MVT=H	7	63
23	SingleParentF=H MUHouse=H ==> MVT=H	5	63
24	SingleParentF=H MUHouse=H Income=L ==> MVT=H	5	63
25	SingleParentF=H HomeOwn=L HeterogeneityInx=H ==> MVT=H	5	63
26	SingleParentF=H HomeOwn=L HeterogeneityInx=H Income=L ==> MVT=H	5	63
27	Male17-24=H UnstableRent=H MUHouse=H Income=L ==> MVT=H	5	62
28	Male17-24=H HeterogeneityInx=H Employment=M Income=L ==> MVT=H	5	62
29	HomeOwn=L Male17-24=H MUHouse=H ==> MVT=H	6	62
30	Male17-24=H MUHouse=H Income=L ==> MVT=H	6	61
31	Male17-24=H HeterogeneityInx=H Employment=M ==> MVT=H	5	60
32	HomeOwn=L UnstableRent=H HEdu=M MUHouse=H ==> MVT=H	5	60
33	HomeOwn=L HeterogeneityInx=H Income=L ==> MVT=H	7	60
34	HomeOwn=L HeterogeneityInx=H HEdu=M ==> MVT=H	5	60

APPENDIX B: (continued)

35	HomeOwn=L Male17-24=H UnstableRent=H MUHouse=H ==> MVT=H	5	60
36	Male17-24=H UnstableRent=H HEdu=M Income=L ==> MVT=H	5	59
37	HomeOwn=L HeterogeneityInx=H ==> MVT=H	8	59
38	HEdu=M MUHouse=H Income=L ==> MVT=H	5	58
39	UnstableRent=H HEdu=M MUHouse=H ==> MVT=H	6	58
40	Male17-24=H UnstableRent=H HEdu=M ==> MVT=H	6	57
41	HomeOwn=L Male17-24=H HEdu=M ==> MVT=H	5	57
42	SingleParentF=H Male17-24=H HeterogeneityInx=H Income=L ==> MVT=H	6	57
43	UnstableRent=H MUHouse=H Income=L ==> MVT=H	7	56
44	Male17-24=H MUHouse=H ==> MVT=H	7	56
45	Male17-24=H UnstableRent=H MUHouse=H ==> MVT=H	6	56
46	HomeOwn=L UnstableRent=H MUHouse=H Income=L ==> MVT=H	6	56
47	HomeOwn=L HEdu=M MUHouse=H ==> MVT=H	5	55
48	SingleParentF=H Male17-24=H HeterogeneityInx=H ==> MVT=H	7	55
49	SingleParentF=H HomeOwn=L HEdu=M ==> MVT=H	6	55
50	SingleParentF=H HomeOwn=L HEdu=M Income=L ==> MVT=H	5	55
51	MUHouse=H Income=L ==> MVT=H	8	54
52	HomeOwn=L MUHouse=H Income=L ==> MVT=H	6	54
53	HomeOwn=L Male17-24=H ==> MVT=H	8	53
54	HomeOwn=L Male17-24=H UnstableRent=H ==> MVT=H	6	53
55	HomeOwn=L Male17-24=H UnstableRent=H Income=L ==> MVT=H	5	53
56	SingleParentF=H HomeOwn=L ==> MVT=H	8	53
57	SingleParentF=H HomeOwn=L Male17-24=H ==> MVT=H	5	53
58	HEdu=M MUHouse=H ==> MVT=H	7	52
59	HomeOwn=L Male17-24=H Income=L ==> MVT=H	7	52
60	SingleParentF=H HomeOwn=L Male17-24=H Income=L ==> MVT=H	5	52
61	Male17-24=H UnstableRent=H Income=L ==> MVT=H	6	51
62	HomeOwn=L UnstableRent=H MUHouse=H ==> MVT=H	8	51
63	SingleParentF=H HomeOwn=L Income=L ==> MVT=H	7	51
64	SingleParentF=H HomeOwn=L UnstableRent=H ==> MVT=H	5	51
65	SingleParentF=H HomeOwn=L UnstableRent=H Income=L ==> MVT=H	5	51
66	HomeOwn=L UnstableRent=H HEdu=M Income=L ==> MVT=H	5	50
67	SingleParentF=H HomeOwn=L AfricanA=H ==> MVT=H	6	50
68	Male17-24=H Employment=M Income=L ==> MVT=H	6	49
69	Male17-24=H HeterogeneityInx=H ==> MVT=H	9	49
70	Male17-24=H HeterogeneityInx=H Income=L ==> MVT=H	8	49
71	Male17-24=H HeterogeneityInx=H HEdu=M ==> MVT=H	6	49
72	HomeOwn=L MUHouse=H ==> MVT=H	9	49
73	HomeOwn=L UnstableRent=H HEdu=M ==> MVT=H	6	49
74	SingleParentF=H HomeOwn=L AfricanA=H Income=L ==> MVT=H	6	49
75	UnstableRent=H HEdu=M Income=L ==> MVT=H	6	48

APPENDIX B: (continued)

76	Male17-24=H UnstableRent=H ==> MVT=H	7	48
77	HomeOwn=L HEdu=M Income=L ==> MVT=H	7	48
78	UnstableRent=H MUHouse=H ==> MVT=H	9	47
79	UnstableRent=H HEdu=M ==> MVT=H	7	47
80	Male17-24=H Employment=M ==> MVT=H	6	47
81	HomeOwn=L UnstableRent=H Income=L ==> MVT=H	7	47
82	HomeOwn=L AfricanA=H ==> MVT=H	7	47
83	Employment=H UnstableRent=H ==> MVT=H	5	46
84	UnstableRent=H Income=L ==> MVT=H	9	45
85	Male17-24=H HeterogeneityInx=H AfricanA=H ==> MVT=H	5	45
86	HomeOwn=L UnstableRent=H ==> MVT=H	10	45
87	HomeOwn=L AfricanA=H Income=L ==> MVT=H	6	45
88	SingleParentF=H UnstableRent=H ==> MVT=H	6	45
89	SingleParentF=H UnstableRent=H Income=L ==> MVT=H	6	45
90	SingleParentF=H HeterogeneityInx=H Income=L ==> MVT=H	8	45
91	SingleParentF=H HeterogeneityInx=H HEdu=M ==> MVT=H	6	45
92	SingleParentF=H HeterogeneityInx=H HEdu=M Income=L ==> MVT=H	5	45
93	HomeOwn=L Income=L ==> MVT=H	11	44
94	SingleParentF=H HeterogeneityInx=H ==> MVT=H	8	44
95	MUHouse=H ==> MVT=H	11	43
96	HeterogeneityInx=H Employment=M Income=L ==> MVT=H	6	43
97	SingleParentF=H HeterogeneityInx=H AfricanA=H ==> MVT=H	6	43
98	SingleParentF=H HeterogeneityInx=H AfricanA=H Income=L ==> MVT=H	5	43
99	Employment=H MUHouse=H ==> MVT=H	6	42
100	HomeOwn=L Employment=H ==> MVT=H	6	42
101	SingleParentF=H Male17-24=H ==> MVT=H	8	42
102	SingleParentF=H Male17-24=H Income=L ==> MVT=H	8	42
103	SingleParentF=H Male17-24=H HEdu=M ==> MVT=H	5	42
104	HomeOwn=L HEdu=M ==> MVT=H	8	41
105	UnstableRent=H ==> MVT=H	11	40
106	HeterogeneityInx=H AfricanA=H Income=L ==> MVT=H	6	40
107	Male17-24=H Income=L ==> MVT=H	11	40
108	Male17-24=H HEdu=M Income=L ==> MVT=H	7	40

APPENDIX C: REPRESENTATIVE MINED SARS FOR DANGEROUS STREET
DUE TO THEFTS FROM MOTOR VEHICLE USING CRISP MAPPING, SUPPORT
THRESHOLD OF 5%, CONFIDENCE THRESHOLD OF 40%

	Rules	S%	C%
1	HomeOwn=L Male17-24=H HeterogeneityInx=H MUHouse=H ==> TFM=H	5	69
2	Male17-24=H HEdu=M MUHouse=H ==> TFM=H	5	67
3	HomeOwn=L UnstableRent=H HEdu=M MUHouse=H ==> TFM=H HomeOwn=L HeterogeneityInx=H UnstableRent=H MUHouse=H ==>	6	67
4	TFM=H	5	67
5	HomeOwn=L HeterogeneityInx=H MUHouse=H ==> TFM=H	6	64
6	HomeOwn=L Male17-24=H MUHouse=H ==> TFM=H	6	64
7	UnstableRent=H HEdu=M MUHouse=H ==> TFM=H	6	63
8	HomeOwn=L Employment=H UnstableRent=H MUHouse=H ==> TFM=H	5	63
9	HomeOwn=L Male17-24=H UnstableRent=H MUHouse=H ==> TFM=H	6	63
10	HomeOwn=L Male17-24=H UnstableRent=H HEdu=M ==> TFM=H	5	63
11	HomeOwn=L Male17-24=H HeterogeneityInx=H ==> TFM=H	6	63
12	SingleParentF=H MUHouse=H ==> TFM=H	5	63
13	HeterogeneityInx=H UnstableRent=H MUHouse=H ==> TFM=H	6	61
14	HomeOwn=L HEdu=M MUHouse=H ==> TFM=H	6	60
15	HomeOwn=L Employment=L Income=L ==> TFM=MH	6	60
16	Male17-24=H HeterogeneityInx=H MUHouse=H ==> TFM=H	5	59
17	HomeOwn=L UnstableRent=H MUHouse=H ==> TFM=H	9	59
18	HomeOwn=L HeterogeneityInx=H UnstableRent=H ==> TFM=H	6	59
19	AfricanA=H MUHouse=M ==> TFM=MH	6	59
20	AfricanA=H Employment=L Income=L ==> TFM=MH	6	59
21	AfricanA=H MUHouse=M Income=L ==> TFM=MH	6	58
22	AfricanA=H Employment=L ==> TFM=MH	6	58
23	HeterogeneityInx=H MUHouse=H ==> TFM=H	7	57
24	Male17-24=H MUHouse=H Income=L ==> TFM=H	5	57
25	Male17-24=H UnstableRent=H HEdu=M ==> TFM=H	6	57
26	HomeOwn=L Employment=H MUHouse=H ==> TFM=H	5	57
27	HomeOwn=L Employment=H UnstableRent=H ==> TFM=H	5	57
28	HomeOwn=L Male17-24=H HEdu=M ==> TFM=H	5	57
29	SingleParentF=H HomeOwn=L Male17-24=H ==> TFM=H	6	57
30	Employment=L Income=L ==> TFM=MH	7	57
31	Male17-24=H MUHouse=M Income=L ==> TFM=MH	5	57
32	Male17-24=L HEdu=M Income=L ==> TFM=MH	6	57
33	HomeOwn=L MUHouse=M ==> TFM=MH	6	57
34	HomeOwn=L Employment=L ==> TFM=MH	6	57
35	HEdu=M MUHouse=H ==> TFM=H	7	56

APPENDIX C: (continued)

36	Employment=H UnstableRent=H MUHouse=H ==> TFM=H	6	56
37	HeterogeneityInx=H MUHouse=H Income=L ==> TFM=H	5	56
38	HeterogeneityInx=L Income=L ==> TFM=MH	6	56
39	Male17-24=L MUHouse=L Income=L ==> TFM=MH	5	56
40	SingleParentF=L Income=L ==> TFM=MH	5	56
41	Male17-24=H HeterogeneityInx=H UnstableRent=H ==> TFM=H	5	55
42	HomeOwn=L MUHouse=H ==> TFM=H	10	55
43	SingleParentF=H Male17-24=H UnstableRent=H ==> TFM=H	5	55
44	SingleParentF=H Male17-24=H UnstableRent=H Income=L ==> TFM=H	5	55
45	MUHouse=M Income=L ==> TFM=MH	9	55
46	HomeOwn=MH Income=L ==> TFM=MH	7	55
47	HomeOwn=MH UnstableRent=M ==> TFM=MH	7	55
48	HeterogeneityInx=H UnstableRent=H ==> TFM=H	6	54
49	Male17-24=H MUHouse=H ==> TFM=H	7	54
50	Male17-24=H UnstableRent=H MUHouse=H ==> TFM=H	6	54
51	HomeOwn=L UnstableRent=H MUHouse=H Income=L ==> TFM=H	5	54
52	SingleParentF=H HomeOwn=L Male17-24=H Income=L ==> TFM=H	5	54
53	UnstableRent=L MUHouse=L Income=L ==> TFM=MH	7	54
54	Employment=L ==> TFM=MH	8	54
55	Male17-24=L Income=L ==> TFM=MH	9	54
56	SingleParentF=MH Income=L ==> TFM=MH	7	54
57	UnstableRent=H MUHouse=H ==> TFM=H	10	53
58	UnstableRent=H MUHouse=H Income=L ==> TFM=H	6	53
59	HomeOwn=L Male17-24=H UnstableRent=H ==> TFM=H	6	53
60	UnstableRent=L Income=L ==> TFM=MH	7	53
61	Male17-24=L HeterogeneityInx=MH ==> TFM=MH	5	53
62	SingleParentF=H MUHouse=M Income=L ==> TFM=MH	5	53
63	SingleParentF=MH ==> TFM=MH	9	53
64	HomeOwn=L HeterogeneityInx=H ==> TFM=H	7	52
65	HEdu=M MUHouse=M Income=L ==> TFM=MH	5	52
66	Male17-24=H MUHouse=M ==> TFM=MH	5	52
67	HomeOwn=MH MUHouse=M ==> TFM=MH	6	52
68	HomeOwn=MH HEdu=M ==> TFM=MH	7	52
69	HomeOwn=MH Employment=H ==> TFM=MH	5	52
70	HomeOwn=MH AfircanA=L ==> TFM=MH	7	52
71	SingleParentF=L HomeOwn=MH ==> TFM=MH	5	52
72	HomeOwn=L UnstableRent=H HEdu=M ==> TFM=H	7	51
73	HomeOwn=L Male17-24=H ==> TFM=H	8	51
74	HomeOwn=L Male17-24=H UnstableRent=H Income=L ==> TFM=H	5	51
75	HEdu=M MUHouse=L Income=L ==> TFM=MH	6	51
76	UnstableRent=M HEdu=L ==> TFM=MH	6	51

APPENDIX C: (continued)

77	Employment=H Income=L ==> TFM=MH	8	51
78	HomeOwn=MH ==> TFM=MH	12	51
79	Employment=H UnstableRent=H ==> TFM=H	6	50
80	HomeOwn=L MUHouse=H Income=L ==> TFM=H	6	50
81	SingleParentF=H UnstableRent=H HEdu=M ==> TFM=H	5	50
82	UnstableRent=M Income=L ==> TFM=MH	9	50
83	UnstableRent=M MUHouse=M ==> TFM=MH	9	50
84	UnstableRent=M HEdu=M Income=L ==> TFM=MH	5	50
85	Employment=H MUHouse=M ==> TFM=MH	7	50
86	AfricanA=L Employment=H MUHouse=M ==> TFM=MH	5	50
87	AfricanA=L Employment=H UnstableRent=M ==> TFM=MH	5	50
88	HeterogeneityInx=H Employment=H Income=L ==> TFM=MH	5	50
89	HeterogeneityInx=MH HEdu=M ==> TFM=MH	7	50
90	Male17-24=H UnstableRent=M ==> TFM=MH	5	50
91	HomeOwn=MH Employment=M ==> TFM=MH	5	50
92	SingleParentF=H MUHouse=M ==> TFM=MH	5	50
93	MUHouse=H Income=L ==> TFM=H	7	49
94	SingleParentF=H Male17-24=H HeterogeneityInx=H Income=L ==> TFM=H	5	49
95	SingleParentF=H HomeOwn=L UnstableRent=H ==> TFM=H	5	49
96	MUHouse=L Income=L ==> TFM=MH	10	49
97	Employment=H UnstableRent=M ==> TFM=MH	7	49
98	HeterogeneityInx=MH Income=L ==> TFM=MH	6	49
99	HeterogeneityInx=L UnstableRent=M ==> TFM=MH	6	49
100	HomeOwn=MH Male17-24=L ==> TFM=MH	6	49
101	UnstableRent=H HEdu=M ==> TFM=H	8	48
102	Employment=H MUHouse=H ==> TFM=H	7	48
103	HomeOwn=L UnstableRent=H ==> TFM=H	10	48
104	HomeOwn=L Employment=H ==> TFM=H	7	48
105	HomeOwn=L HeterogeneityInx=H Income=L ==> TFM=H	5	48
106	HomeOwn=L Male17-24=H Income=L ==> TFM=H	6	48
107	SingleParentF=H Male17-24=H HeterogeneityInx=H ==> TFM=H	6	48
108	UnstableRent=M ==> TFM=MH	15	48
109	Employment=H HEdu=M Income=L ==> TFM=MH	5	48
110	Employment=H HEdu=M MUHouse=L ==> TFM=MH	5	48
111	AfricanA=MH ==> TFM=MH	7	48
112	Male17-24=L AfricanA=L UnstableRent=M ==> TFM=MH	6	48
113	SingleParentF=L Male17-24=L HeterogeneityInx=L HEdu=L ==> TFM=MH	6	48
114	Male17-24=H UnstableRent=H Income=L ==> TFM=H	6	47
115	MUHouse=M ==> TFM=MH	15	47
116	AfricanA=H HEdu=M ==> TFM=MH	10	47
117	AfricanA=H HEdu=M Income=L ==> TFM=MH	9	47

APPENDIX C: (continued)

118	Male17-24=L UnstableRent=M ==> TFM=MH	8	47
119	Male17-24=L HeterogeneityInx=L HEdu=L ==> TFM=MH	7	47
120	Male17-24=L HeterogeneityInx=L AfircanA=L HEdu=L ==> TFM=MH	6	47
121	SingleParentF=L Employment=H HEdu=L ==> TFM=MH	6	47
122	SingleParentF=L HeterogeneityInx=L HEdu=L ==> TFM=MH	7	47
123	MUHouse=H ==> TFM=H	12	46
124	Male17-24=H UnstableRent=H ==> TFM=H	7	46
125	SingleParentF=H UnstableRent=H ==> TFM=H	6	46
126	SingleParentF=H Male17-24=H HEdu=M ==> TFM=H	5	46
127	SingleParentF=H HomeOwn=L ==> TFM=H	7	46
128	Income=L ==> TFM=MH	23	46
129	HEdu=M Income=L ==> TFM=MH	15	46
130	AfricanA=H Income=L ==> TFM=MH	13	46
131	HeterogeneityInx=L MUHouse=M ==> TFM=MH	7	46
132	HeterogeneityInx=L HEdu=L ==> TFM=MH	8	46
133	HeterogeneityInx=L AfircanA=L HEdu=L ==> TFM=MH	7	46
134	Male17-24=L Employment=H HEdu=L ==> TFM=MH	5	46
135	HomeOwn=L Male17-24=L ==> TFM=MH	8	46
136	SingleParentF=L AfircanA=L Employment=H HEdu=M ==> TFM=MH	6	46
137	SingleParentF=L HeterogeneityInx=L AfircanA=L HEdu=L ==> TFM=MH	7	46
138	SingleParentF=L Male17-24=L UnstableRent=M ==> TFM=MH	5	46
139	Male17-24=H HeterogeneityInx=H HEdu=M ==> TFM=H	5	45
140	SingleParentF=H UnstableRent=H Income=L ==> TFM=H	6	45
141	SingleParentF=H Male17-24=H HEdu=M Income=L ==> TFM=H	5	45
142	Employment=H HEdu=L ==> TFM=MH	7	45
143	Employment=H UnstableRent=L HEdu=M ==> TFM=MH	5	45
144	AfricanA=H ==> TFM=MH	14	45
145	AfricanA=H MUHouse=L Income=L ==> TFM=MH	5	45
146	AfircanA=L UnstableRent=M MUHouse=M ==> TFM=MH	5	45
147	AfircanA=L Employment=H HEdu=M ==> TFM=MH	7	45
148	HeterogeneityInx=H MUHouse=L Income=L ==> TFM=MH	5	45
149	HeterogeneityInx=H UnstableRent=L MUHouse=L ==> TFM=MH	5	45
150	HeterogeneityInx=MH ==> TFM=MH	10	45
151	Male17-24=L UnstableRent=M MUHouse=M ==> TFM=MH	5	45
152	Male17-24=L AfircanA=L Employment=H ==> TFM=MH	9	45
153	HomeOwn=L AfricanA=H Income=L ==> TFM=MH	6	45
154	SingleParentF=L UnstableRent=M ==> TFM=MH	7	45
155	SingleParentF=L AfircanA=L UnstableRent=M ==> TFM=MH	6	45
156	SingleParentF=L Male17-24=L Employment=H UnstableRent=L ==> TFM=MH	5	45
157	HEdu=M MUHouse=M ==> TFM=MH	8	44

APPENDIX C: (continued)

158	HEdu=L Income=L ==> TFM=MH	7	44
159	UnstableRent=M HEdu=M ==> TFM=MH	8	44
160	AfricanA=L Income=L ==> TFM=MH	6	44
161	AfricanA=L UnstableRent=M ==> TFM=MH	8	44
162	HeterogeneityInx=H UnstableRent=L ==> TFM=MH	6	44
163	HeterogeneityInx=MH AfricanA=L ==> TFM=MH	5	44
164	Male17-24=L MUHouse=M ==> TFM=MH	8	44
165	Male17-24=L Employment=H ==> TFM=MH	12	44
166	HomeOwn=H Income=L ==> TFM=MH	6	44
167	HomeOwn=L AfricanA=H ==> TFM=MH	6	44
168	SingleParentF=H AfricanA=H HEdu=M ==> TFM=MH	7	44
169	SingleParentF=H AfricanA=H HEdu=M Income=L ==> TFM=MH	7	44
170	SingleParentF=H HomeOwn=L AfricanA=H Income=L ==> TFM=MH	5	44
171	SingleParentF=L HEdu=L ==> TFM=MH	9	44
172	SingleParentF=L Employment=H UnstableRent=L ==> TFM=MH	6	44
173	SingleParentF=L Employment=H UnstableRent=L MUHouse=L ==> TFM=MH	5	44
174	SingleParentF=L Male17-24=L HEdu=L ==> TFM=MH	8	44
175	SingleParentF=L Male17-24=L Employment=H ==> TFM=MH	9	44
176	SingleParentF=L Male17-24=L AfricanA=L Employment=H ==> TFM=MH	8	44
177	SingleParentF=L Male17-24=L HeterogeneityInx=L Employment=H ==> TFM=MH	6	44
178	SingleParentF=L HomeOwn=H Employment=H UnstableRent=L ==> TFM=MH	5	44
179	UnstableRent=H ==> TFM=H	12	43
180	SingleParentF=H HomeOwn=L Income=L ==> TFM=H	6	43
181	Employment=H UnstableRent=L ==> TFM=MH	8	43
182	AfricanA=L Employment=H ==> TFM=MH	13	43
183	AfricanA=L Employment=H HEdu=L ==> TFM=MH	5	43
184	HeterogeneityInx=H Employment=H ==> TFM=MH	8	43
185	Male17-24=H Income=L ==> TFM=MH	11	43
186	Male17-24=H Employment=H ==> TFM=MH	6	43
187	Male17-24=L HEdu=L ==> TFM=MH	10	43
188	Male17-24=L Employment=H HEdu=M ==> TFM=MH	6	43
189	Male17-24=L AfricanA=L HEdu=L ==> TFM=MH	8	43
190	Male17-24=L HeterogeneityInx=L Employment=H ==> TFM=MH	6	43
191	Male17-24=L HeterogeneityInx=L AfricanA=L Employment=H ==> TFM=MH	5	43
192	SingleParentF=H HomeOwn=L AfricanA=H ==> TFM=MH	5	43
193	SingleParentF=L Employment=H ==> TFM=MH	13	43
194	SingleParentF=L AfricanA=L Employment=H ==> TFM=MH	11	43
195	SingleParentF=L AfricanA=L Employment=H UnstableRent=L ==> TFM=MH	5	43

APPENDIX C: (continued)

196	AfricanA=L MUHouse=H ==> TFM=H	6	42
197	Male17-24=H Employment=M Income=L ==> TFM=H	5	42
198	SingleParentF=H Male17-24=H ==> TFM=H	8	42
199	SingleParentF=H Male17-24=H Income=L ==> TFM=H	8	42
200	HEdu=M MUHouse=L ==> TFM=MH	10	42
201	Employment=H ==> TFM=MH	19	42
202	Employment=H MUHouse=L ==> TFM=MH	7	42
203	Employment=H UnstableRent=L MUHouse=L ==> TFM=MH	7	42
204	AfricanA=H MUHouse=L ==> TFM=MH	6	42
205	Male17-24=H AfricanA=H ==> TFM=MH	8	42
206	Male17-24=H AfricanA=H Income=L ==> TFM=MH	7	42
207	Male17-24=H HeterogeneityInx=H Income=L ==> TFM=MH	7	42
208	Male17-24=L Employment=H MUHouse=L ==> TFM=MH	5	42
209	Male17-24=L Employment=H UnstableRent=L ==> TFM=MH	6	42
210	HomeOwn=H Employment=H MUHouse=L ==> TFM=MH	6	42
211	HomeOwn=H Employment=H UnstableRent=L MUHouse=L ==> TFM=MH	6	42
212	HomeOwn=L Income=L ==> TFM=MH	10	42
213	HomeOwn=L HEdu=M Income=L ==> TFM=MH	6	42
214	HomeOwn=L HEdu=L ==> TFM=MH	5	42
215	HomeOwn=L HeterogeneityInx=L ==> TFM=MH	6	42
216	SingleParentF=H AfricanA=H Income=L ==> TFM=MH	10	42
217	SingleParentF=L MUHouse=M ==> TFM=MH	7	42
218	SingleParentF=L Employment=H MUHouse=L ==> TFM=MH	5	42
219	SingleParentF=L AfricanA=L HEdu=L ==> TFM=MH	8	42
220	SingleParentF=L Male17-24=L AfricanA=L HEdu=L ==> TFM=MH	7	42
221	SingleParentF=L HomeOwn=H Employment=H ==> TFM=MH	6	42
222	UnstableRent=H HEdu=M Income=L ==> TFM=H	5	41
223	Male17-24=H Employment=M ==> TFM=H	5	41
224	HomeOwn=L UnstableRent=H Income=L ==> TFM=H	7	41
225	HomeOwn=L AfricanA=L ==> TFM=H	6	41
226	HEdu=L ==> TFM=MH	15	41
227	UnstableRent=L HEdu=M ==> TFM=MH	9	41
228	Employment=H HEdu=M ==> TFM=MH	10	41
229	Employment=M UnstableRent=M ==> TFM=MH	5	41
230	AfricanA=L MUHouse=M ==> TFM=MH	7	41
231	AfricanA=L HEdu=L ==> TFM=MH	9	41
232	HeterogeneityInx=H MUHouse=L ==> TFM=MH	7	41
233	HeterogeneityInx=H UnstableRent=M ==> TFM=MH	5	41
234	HeterogeneityInx=H AfricanA=H Income=L ==> TFM=MH	6	41
235	Male17-24=H ==> TFM=MH	14	41
236	Male17-24=H HeterogeneityInx=H ==> TFM=MH	8	41

APPENDIX C: (continued)

237	Male17-24=L ==> TFM=MH	23	41
238	Male17-24=L HEdu=M MUHouse=L ==> TFM=MH	7	41
239	Male17-24=L UnstableRent=L HEdu=M ==> TFM=MH	7	41
240	Male17-24=L UnstableRent=L HEdu=M MUHouse=L ==> TFM=MH	6	41
241	HomeOwn=H Employment=H UnstableRent=L ==> TFM=MH	6	41
242	HomeOwn=H Male17-24=L HEdu=M MUHouse=L ==> TFM=MH	5	41
243	SingleParentF=H HEdu=M Income=L ==> TFM=MH	8	41
244	SingleParentF=H AfricanA=H ==> TFM=MH	10	41
245	SingleParentF=L Employment=H HEdu=M ==> TFM=MH	6	41
246	SingleParentF=L HomeOwn=H AfricanA=L Employment=H ==> TFM=MH	5	41
247	UnstableRent=H Income=L ==> TFM=H	8	40
248	Male17-24=H HeterogeneityInx=H Income=L ==> TFM=H	7	40
249	MUHouse=L ==> TFM=MH	17	40
250	HEdu=M ==> TFM=MH	23	40
251	UnstableRent=L HEdu=M MUHouse=L ==> TFM=MH	8	40
252	AfricanA=L HEdu=M MUHouse=L ==> TFM=MH	5	40
253	AfricanA=L Employment=H UnstableRent=L ==> TFM=MH	5	40
254	HeterogeneityInx=H Income=L ==> TFM=MH	11	40
255	HeterogeneityInx=L ==> TFM=MH	16	40
256	Male17-24=L HEdu=M ==> TFM=MH	13	40
257	Male17-24=L AfricanA=L MUHouse=M ==> TFM=MH	5	40
258	Male17-24=L HeterogeneityInx=L ==> TFM=MH	13	40
259	HomeOwn=H HEdu=M MUHouse=L ==> TFM=MH	7	40
260	HomeOwn=H UnstableRent=L HEdu=M MUHouse=L ==> TFM=MH	6	40
261	HomeOwn=H Employment=H ==> TFM=MH	8	40
262	HomeOwn=H Employment=H HEdu=M ==> TFM=MH	5	40
263	SingleParentF=H MUHouse=L ==> TFM=MH	5	40
264	SingleParentF=H HEdu=M ==> TFM=MH	8	40
265	SingleParentF=L HEdu=M MUHouse=L ==> TFM=MH	5	40
266	SingleParentF=L HeterogeneityInx=L Employment=H ==> TFM=MH	6	40
267	SingleParentF=L HeterogeneityInx=L AfricanA=L Employment=H ==> TFM=MH	6	40
268	SingleParentF=L Male17-24=L MUHouse=M ==> TFM=MH	5	40
269	SingleParentF=L HomeOwn=H MUHouse=L ==> TFM=MH	7	40
270	SingleParentF=L HomeOwn=H UnstableRent=L MUHouse=L ==> TFM=MH	7	40
271	SingleParentF=L HomeOwn=H Male17-24=L Employment=H ==> TFM=MH	5	40

APPENDIX D: REPRESENTATIVE MINED FUZZY SARS FOR DANGEROUS
STREET DUE TO CRIME OF ALL TYPES USING SUPPORT THRESHOLD OF 5%,
CONFIDENCE THRESHOLD OF 40%

Rules	S%	C%
1 Male17-24_H UnstableRent_M ==> CAT_MH	5	77
2 HomeOwn_MH Employment_M UnstableRent_M ==> CAT_MH	6	77
3 HomeOwn_MH MUHouse_M Income_L ==> CAT_MH	6	75
4 SingleParentF_H UnstableRent_H MUHouse_H ==> CAT_H	5	75
5 HomeOwn_MH UnstableRent_M HEdu_L ==> CAT_MH	5	73
6 HomeOwn_MH HeterogeneityInx_MH MUHouse_M ==> CAT_MH	6	73
7 HeterogeneityInx_MH MUHouse_M Income_L ==> CAT_MH	5	73
8 SingleParentF_H HomeOwn_L HEdu_M Income_L ==> CAT_H	5	73
9 SingleParentF_H MUHouse_H Income_L ==> CAT_H	5	72
10 HomeOwn_MH HeterogeneityInx_MH UnstableRent_M ==> CAT_MH	6	72
11 SingleParentF_H MUHouse_H ==> CAT_H	6	72
12 SingleParentF_H HomeOwn_L HEdu_M ==> CAT_H	5	72
13 HomeOwn_MH Male17-24_H ==> CAT_MH	7	71
14 HomeOwn_MH Male17-24_H Income_L ==> CAT_MH	5	70
15 HomeOwn_L HeterogeneityInx_H Income_L ==> CAT_H	5	70
16 SingleParentF_H UnstableRent_M ==> CAT_MH	5	69
17 HomeOwn_MH UnstableRent_M Income_L ==> CAT_MH	7	69
18 HomeOwn_MH UnstableRent_M MUHouse_M ==> CAT_MH	8	69
19 SingleParentF_H UnstableRent_M Income_L ==> CAT_MH	5	69
20 SingleParentF_H MUHouse_L Income_L ==> CAT_MH	7	68
21 SingleParentF_H HomeOwn_MH Income_L ==> CAT_MH	7	68
22 SingleParentF_H HomeOwn_MH ==> CAT_MH	7	68
23 HomeOwn_MH HeterogeneityInx_MH Income_L ==> CAT_MH	7	68
24 Male17-24_H MUHouse_M ==> CAT_MH	5	68
25 HomeOwn_MH UnstableRent_M ==> CAT_MH	13	68
26 HeterogeneityInx_H UnstableRent_M Income_L ==> CAT_MH	5	67
27 HomeOwn_MH AfricanA_M ==> CAT_MH	7	67
28 AfricanA_H UnstableRent_M ==> CAT_MH	5	67
29 HeterogeneityInx_H UnstableRent_H Income_L ==> CAT_H	5	67
30 HeterogeneityInx_MH AfricanA_M ==> CAT_MH	6	67
31 SingleParentF_H HomeOwn_MH AfricanA_H Income_L ==> CAT_MH	5	67

APPENDIX D: (continued)

32	Employment_M UnstableRent_M Income_L ==> CAT_MH	5	67
33	SingleParentF_H MUHouse_L ==> CAT_MH	7	66
34	SingleParentF_H HomeOwn_MH AfricanA_H ==> CAT_MH	5	66
35	HomeOwn_L HeterogeneityInx_H ==> CAT_H	6	66
36	UnstableRent_M MUHouse_M Income_L ==> CAT_MH	5	66
37	HeterogeneityInx_MH UnstableRent_M ==> CAT_MH	8	66
38	Male17-24_MH AfricanA_M ==> CAT_MH	5	66
39	SingleParentF_H HomeOwn_L UnstableRent_H ==> CAT_H	5	66
40	HeterogeneityInx_MH MUHouse_M ==> CAT_MH	9	66
41	HomeOwn_MH HeterogeneityInx_MH ==> CAT_MH	11	66
42	HomeOwn_MH AfricanA_H Income_L ==> CAT_MH	6	65
43	Employment_M UnstableRent_M ==> CAT_MH	8	65
44	SingleParentF_H AfricanA_H MUHouse_L ==> CAT_MH	5	65
45	Male17-24_MH UnstableRent_M ==> CAT_MH	7	65
46	HomeOwn_MH HeterogeneityInx_MH HEdu_M ==> CAT_MH	6	65
47	HeterogeneityInx_MH HEdu_L ==> CAT_MH	7	65
48	SingleParentF_H MUHouse_M ==> CAT_MH	5	65
49	SingleParentF_L HomeOwn_MH AfricanA_L UnstableRent_M ==> CAT_MH	5	64
50	UnstableRent_M MUHouse_L ==> CAT_MH	5	64
51	HomeOwn_MH MUHouse_L Income_L ==> CAT_MH	5	64
52	SingleParentF_MH UnstableRent_M ==> CAT_MH	7	64
53	SingleParentF_H MUHouse_M Income_L ==> CAT_MH	5	64
54	HeterogeneityInx_H UnstableRent_M ==> CAT_MH	7	64
55	HomeOwn_MH MUHouse_M ==> CAT_MH	11	64
56	HomeOwn_MH HEdu_L ==> CAT_MH	10	64
57	SingleParentF_H HomeOwn_L ==> CAT_H	7	63
58	HomeOwn_L UnstableRent_H MUHouse_H Income_L ==> CAT_H	5	63
59	HomeOwn_MH HeterogeneityInx_H Income_L ==> CAT_MH	5	63
60	HomeOwn_MH AfricanA_H ==> CAT_MH	7	63
61	HomeOwn_MH Employment_M Income_L ==> CAT_MH	7	63
62	SingleParentF_H HomeOwn_L Income_L ==> CAT_H	6	63
63	HomeOwn_L MUHouse_H Income_L ==> CAT_H	6	63
64	HomeOwn_MH Male17-24_MH Income_L ==> CAT_MH	6	63
65	HeterogeneityInx_H UnstableRent_H MUHouse_H ==> CAT_H	5	63
66	AfricanA_H MUHouse_M Income_L ==> CAT_MH	5	63
67	UnstableRent_M HEdu_L ==> CAT_MH	7	63

APPENDIX D: (continued)

68	HomeOwn_MH Income_L ==> CAT_MH	14	63
69	HomeOwn_MH Employment_M ==> CAT_MH	10	63
70	Male17-24_H HeterogeneityInx_MH ==> CAT_MH	6	62
71	Male17-24_MH HEdu_L ==> CAT_MH	5	62
72	HomeOwn_MH HeterogeneityInx_H ==> CAT_MH	7	62
73	SingleParentF_L HomeOwn_MH UnstableRent_M ==> CAT_MH	5	62
74	AfricanA_M ==> CAT_MH	13	62
75	MUHouse_M Income_L ==> CAT_MH	10	62
76	UnstableRent_M Income_L ==> CAT_MH	11	62
77	HomeOwn_MH AfricanA_L UnstableRent_M ==> CAT_MH	7	62
78	Employment_L Income_L ==> CAT_MH	8	62
79	HomeOwn_MH Male17-24_MH ==> CAT_MH	8	62
80	Employment_M MUHouse_L Income_L ==> CAT_MH	5	62
81	HomeOwn_MH UnstableRent_M HEdu_M ==> CAT_MH	7	62
82	AfricanA_M Income_L ==> CAT_MH	9	61
83	Male17-24_MH Employment_M ==> CAT_MH	6	61
84	AfricanA_H MUHouse_L Income_L ==> CAT_MH	6	61
85	AfricanA_M HEdu_M ==> CAT_MH	7	61
86	HeterogeneityInx_MH Income_L ==> CAT_MH	12	61
87	AfricanA_H MUHouse_M ==> CAT_MH	6	61
88	Male17-24_H MUHouse_H Income_L ==> CAT_H	5	61
89	SingleParentF_H UnstableRent_H HEdu_M Income_L ==> CAT_H	5	61
90	UnstableRent_M ==> CAT_MH	19	61
91	HomeOwn_L Male17-24_H Income_L ==> CAT_H	6	61
92	Male17-24_MH MUHouse_M ==> CAT_MH	5	61
93	Male17-24_H UnstableRent_H HEdu_M ==> CAT_H	5	61
94	HomeOwn_L UnstableRent_H HEdu_M Income_L ==> CAT_H	5	60
95	HomeOwn_MH MUHouse_L ==> CAT_MH	7	60
96	HeterogeneityInx_H UnstableRent_H ==> CAT_H	6	60
97	HomeOwn_MH ==> CAT_MH	23	60
98	HeterogeneityInx_MH HEdu_M MUHouse_M ==> CAT_MH	5	60
99	SingleParentF_H UnstableRent_H HEdu_M ==> CAT_H	6	60
100	SingleParentF_H HomeOwn_L AfricanA_H ==> CAT_H	5	60
101	HomeOwn_MH Employment_H UnstableRent_M ==> CAT_MH	6	60
102	SingleParentF_MH HomeOwn_MH ==> CAT_MH	7	60
103	HomeOwn_MH AfricanA_L HEdu_L ==> CAT_MH	5	60
104	UnstableRent_H MUHouse_H Income_L ==> CAT_H	7	59

APPENDIX D: (continued)

105	AfricanA_H Employment_L Income_L ==> CAT_MH	5	59
106	HeterogeneityInx_MH Employment_M ==> CAT_MH	7	59
107	HomeOwn_MH Income_M ==> CAT_MH	7	59
108	HomeOwn_L AfricanA_H ==> CAT_H	6	59
109	HomeOwn_L HEdu_M Income_L ==> CAT_H	6	58
110	SingleParentF_L HomeOwn_MH Employment_H ==> CAT_MH	6	58
111	AfricanA_H Employment_L ==> CAT_MH	5	58
112	UnstableRent_M MUHouse_M ==> CAT_MH	11	58
113	HomeOwn_MH AfricanA_L MUHouse_M ==> CAT_MH	5	58
114	SingleParentF_MH HeterogeneityInx_MH Income_L ==> CAT_MH	5	58
115	HeterogeneityInx_MH HEdu_M Income_L ==> CAT_MH	7	58
116	Male17-24_MH AfricanA_H ==> CAT_MH	5	58
117	HEdu_L Income_L ==> CAT_MH	9	58
118	SingleParentF_MH HeterogeneityInx_MH ==> CAT_MH	7	57
119	HomeOwn_MH HEdu_M Income_L ==> CAT_MH	8	57
120	MUHouse_L Income_L ==> CAT_MH	12	57
121	SingleParentF_MH AfricanA_M ==> CAT_MH	6	57
122	HeterogeneityInx_MH UnstableRent_H ==> CAT_MH	6	57
123	SingleParentF_MH Male17-24_MH ==> CAT_MH	6	57
124	HomeOwn_L AfricanA_H Income_L ==> CAT_H	6	57
125	UnstableRent_M Income_M ==> CAT_MH	6	57
126	SingleParentF_H UnstableRent_H Income_L ==> CAT_H	7	57
127	HomeOwn_MH HEdu_M MUHouse_M ==> CAT_MH	6	57
128	HEdu_M MUHouse_H Income_L ==> CAT_H	5	57
129	HomeOwn_MH Employment_H ==> CAT_MH	10	57
130	SingleParentF_H UnstableRent_H ==> CAT_H	7	57
131	HomeOwn_MH HEdu_M ==> CAT_MH	12	57
132	Male17-24_H UnstableRent_H Income_L ==> CAT_H	6	57
133	HomeOwn_MH AfricanA_L Employment_H ==> CAT_MH	7	56
134	HEdu_L MUHouse_M ==> CAT_MH	6	56
135	HomeOwn_L HeterogeneityInx_MH ==> CAT_MH	6	56
136	Employment_H UnstableRent_M MUHouse_M ==> CAT_MH	5	56
137	Male17-24_H UnstableRent_H MUHouse_H ==> CAT_H	5	56
138	HeterogeneityInx_H MUHouse_L Income_L ==> CAT_MH	6	56
139	Male17-24_MH MUHouse_L ==> CAT_MH	5	56
140	HomeOwn_L Male17-24_H ==> CAT_H	6	56
141	Employment_H UnstableRent_M ==> CAT_MH	8	56

APPENDIX D: (continued)

142	Male17-24_MH Income_L ==> CAT_MH	10	56
143	UnstableRent_M HEdu_M ==> CAT_MH	10	56
144	Male17-24_MH ==> CAT_MH	14	56
145	SingleParentF_MH HEdu_M ==> CAT_MH	8	56
146	SingleParentF_L HomeOwn_MH ==> CAT_MH	9	55
147	SingleParentF_MH HEdu_M Income_L ==> CAT_MH	6	55
148	AfricanA_H MUHouse_L ==> CAT_MH	7	55
149	HeterogeneityInx_MH ==> CAT_MH	19	55
150	HomeOwn_MH AfricanA_L ==> CAT_MH	10	55
151	HomeOwn_MH MUHouse_H ==> CAT_MH	6	55
152	SingleParentF_MH ==> CAT_MH	13	55
153	UnstableRent_M HEdu_M Income_L ==> CAT_MH	6	55
154	SingleParentF_L HomeOwn_MH AfricanA_L ==> CAT_MH	7	55
155	AfricanA_M Employment_H ==> CAT_MH	5	55
156	UnstableRent_H HEdu_M MUHouse_H ==> CAT_H	6	55
157	HeterogeneityInx_H MUHouse_H ==> CAT_H	6	55
158	HEdu_M MUHouse_L Income_L ==> CAT_MH	7	54
159	SingleParentF_H Income_L ==> CAT_MH	14	54
160	HomeOwn_L UnstableRent_H Income_L ==> CAT_H	7	54
161	SingleParentF_H AfricanA_H Income_L ==> CAT_MH	10	54
162	HomeOwn_MH UnstableRent_L ==> CAT_MH	5	54
163	Employment_L ==> CAT_MH	8	54
164	UnstableRent_H HEdu_M Income_L ==> CAT_H	7	54
165	SingleParentF_L AfricanA_L UnstableRent_M ==> CAT_MH	7	54
166	SingleParentF_MH Income_L ==> CAT_MH	10	54
167	Male17-24_MH HeterogeneityInx_MH ==> CAT_MH	7	54
168	SingleParentF_L UnstableRent_M ==> CAT_MH	8	54
169	AfricanA_H Income_L ==> CAT_MH	13	54
170	HomeOwn_L UnstableRent_H HEdu_M ==> CAT_H	6	53
171	SingleParentF_H AfricanA_H ==> CAT_MH	10	53
172	SingleParentF_L MUHouse_H ==> CAT_MH	6	53
173	Male17-24_MH HEdu_M Income_L ==> CAT_MH	5	53
174	SingleParentF_H HeterogeneityInx_H AfricanA_H Income_L ==> CAT_MH	5	53
175	HomeOwn_MH Male17-24_L ==> CAT_MH	9	53
176	HEdu_M MUHouse_M Income_L ==> CAT_MH	6	53
177	SingleParentF_H ==> CAT_MH	14	53

APPENDIX D: (continued)

178	AfricanA_L UnstableRent_M ==> CAT_MH	9	53
179	Income_L ==> CAT_MH	27	53
180	MUHouse_H Income_L ==> CAT_H	8	53
181	UnstableRent_L Income_L ==> CAT_MH	7	53
182	Employment_H MUHouse_H ==> CAT_MH	7	52
183	Employment_M MUHouse_M ==> CAT_MH	7	52
184	AfricanA_H Employment_M Income_L ==> CAT_MH	6	52
185	HomeOwn_L UnstableRent_H MUHouse_H ==> CAT_H	7	52
186	MUHouse_M ==> CAT_MH	17	52
187	Employment_M Income_L ==> CAT_MH	11	52
188	Male17-24_MH HEdu_M ==> CAT_MH	7	52
189	Male17-24_H Income_L ==> CAT_MH	11	52
190	HeterogeneityInx_H AfricanA_H Income_L ==> CAT_MH	6	52
191	HomeOwn_L MUHouse_H ==> CAT_H	7	52
192	SingleParentF_H HeterogeneityInx_H AfricanA_H ==> CAT_MH	5	52
193	UnstableRent_L MUHouse_L Income_L ==> CAT_MH	6	52
194	AfricanA_H Employment_M ==> CAT_MH	6	52
195	HomeOwn_MH Male17-24_L AfricanA_L ==> CAT_MH	6	52
196	Male17-24_L HeterogeneityInx_MH ==> CAT_MH	6	52
197	SingleParentF_MH MUHouse_L ==> CAT_MH	5	52
198	SingleParentF_H Employment_M ==> CAT_MH	7	52
199	SingleParentF_H Employment_M Income_L ==> CAT_MH	7	52
200	SingleParentF_H HeterogeneityInx_H Income_L ==> CAT_MH	8	52
201	Male17-24_H ==> CAT_MH	13	51
202	SingleParentF_L UnstableRent_M MUHouse_M ==> CAT_MH	5	51
203	SingleParentF_H Male17-24_H Income_L ==> CAT_MH	7	51
204	Male17-24_H MUHouse_H ==> CAT_H	5	51
205	HeterogeneityInx_MH HEdu_M ==> CAT_MH	11	51
206	AfricanA_H UnstableRent_H Income_L ==> CAT_H	6	51
207	Employment_H MUHouse_M ==> CAT_MH	7	51
208	Male17-24_H AfricanA_H Income_L ==> CAT_MH	6	51
209	AfricanA_L Employment_H UnstableRent_M ==> CAT_MH	5	51
210	Male17-24_H Employment_H ==> CAT_MH	6	51
211	AfricanA_H ==> CAT_MH	14	51
212	AfricanA_L UnstableRent_M MUHouse_M ==> CAT_MH	6	50
213	HomeOwn_MH HeterogeneityInx_L ==> CAT_MH	6	50
214	AfricanA_H UnstableRent_H ==> CAT_H	6	50

APPENDIX D: (continued)

215	HomeOwn_L Income_L ==> CAT_H	10	50
216	SingleParentF_H Male17-24_H ==> CAT_MH	7	50
217	Male17-24_MH Employment_H ==> CAT_MH	5	50
218	SingleParentF_L HomeOwn_MH Male17-24_L ==> CAT_MH	5	50
219	HeterogeneityInx_MH MUHouse_L ==> CAT_MH	6	50
220	Male17-24_H AfricanA_H ==> CAT_MH	6	50
221	SingleParentF_MH Employment_H ==> CAT_MH	6	50
222	Male17-24_L Income_L ==> CAT_MH	6	50
223	SingleParentF_H HeterogeneityInx_H ==> CAT_MH	8	50
224	HeterogeneityInx_H Income_L ==> CAT_MH	11	50
225	Male17-24_L UnstableRent_M ==> CAT_MH	7	49
226	UnstableRent_M HEdu_M MUHouse_M ==> CAT_MH	5	49
227	Male17-24_H UnstableRent_H ==> CAT_H	6	49
228	Employment_H UnstableRent_H ==> CAT_MH	6	49
229	Employment_M HEdu_L ==> CAT_MH	6	49
230	HeterogeneityInx_H MUHouse_L ==> CAT_MH	8	49
231	HeterogeneityInx_H AfricanA_H ==> CAT_MH	7	49
232	HeterogeneityInx_H Employment_M Income_L ==> CAT_MH	6	48
233	HomeOwn_L UnstableRent_H ==> CAT_H	9	48
234	HEdu_M Income_L ==> CAT_MH	15	48
235	UnstableRent_H Income_L ==> CAT_H	10	48
236	HeterogeneityInx_MH Employment_H ==> CAT_MH	8	48
237	AfricanA_H HEdu_M Income_L ==> CAT_MH	8	48
238	SingleParentF_L HeterogeneityInx_MH ==> CAT_MH	7	48
239	UnstableRent_H HEdu_M ==> CAT_H	8	48
240	SingleParentF_H HEdu_M Income_L ==> CAT_MH	8	48
241	AfricanA_L MUHouse_H ==> CAT_MH	6	47
242	Male17-24_L AfricanA_L UnstableRent_M ==> CAT_MH	5	47
243	Employment_M UnstableRent_H ==> CAT_H	5	47
244	Employment_M HEdu_M Income_L ==> CAT_MH	6	47
245	HomeOwn_L HEdu_M ==> CAT_H	7	47
246	SingleParentF_H AfricanA_H HEdu_M Income_L ==> CAT_MH	6	47
247	HEdu_L ==> CAT_MH	17	47
248	UnstableRent_H ==> CAT_MH	13	46
249	Male17-24_H HEdu_M Income_L ==> CAT_MH	6	46
250	HeterogeneityInx_MH AfricanA_L ==> CAT_MH	8	46
251	SingleParentF_H AfricanA_H HEdu_M ==> CAT_MH	6	46

APPENDIX D: (continued)

252	Employment_H Income_L ==> CAT_MH	8	46
253	HeterogeneityInx_MH Income_M ==> CAT_MH	6	46
254	SingleParentF_H HEdu_M ==> CAT_MH	8	46
255	Employment_H HEdu_L ==> CAT_MH	7	46
256	HEdu_M MUHouse_M ==> CAT_MH	9	46
257	AfricanA_H HEdu_M ==> CAT_MH	9	46
258	UnstableRent_H MUHouse_H ==> CAT_H	9	46
259	HeterogeneityInx_H ==> CAT_MH	15	45
260	HomeOwn_L Employment_M ==> CAT_H	5	45
261	HomeOwn_H HeterogeneityInx_H ==> CAT_MH	6	45
262	HEdu_M MUHouse_H ==> CAT_H	6	45
263	HeterogeneityInx_H Employment_M ==> CAT_MH	7	45
264	UnstableRent_H Income_L ==> CAT_MH	9	44
265	Male17-24_H HeterogeneityInx_H ==> CAT_MH	6	44
266	Male17-24_H HEdu_M ==> CAT_MH	7	44
267	Employment_M ==> CAT_MH	18	44
268	Male17-24_H HeterogeneityInx_H Income_L ==> CAT_MH	6	44
269	MUHouse_H ==> CAT_MH	11	44
270	HomeOwn_L Income_L ==> CAT_MH	8	44
271	Employment_M MUHouse_L ==> CAT_MH	7	43
272	AfricanA_L Employment_H HEdu_L ==> CAT_MH	5	43
273	UnstableRent_H HEdu_M ==> CAT_MH	7	43
274	Employment_H ==> CAT_MH	19	43
275	HomeOwn_L UnstableRent_H ==> CAT_MH	8	42
276	SingleParentF_L MUHouse_M ==> CAT_MH	7	42
277	SingleParentF_H Male17-24_H Income_L ==> CAT_H	5	42
278	HomeOwn_L UnstableRent_H Income_L ==> CAT_MH	5	42
279	HomeOwn_L ==> CAT_MH	12	42
280	AfricanA_L MUHouse_M ==> CAT_MH	7	42
281	HEdu_M MUHouse_H ==> CAT_MH	5	42
282	Male17-24_H HeterogeneityInx_H Income_L ==> CAT_H	5	42
283	HEdu_L MUHouse_L ==> CAT_MH	6	42
284	Male17-24_H UnstableRent_H ==> CAT_MH	5	42
285	HEdu_M ==> CAT_MH	24	41
286	HomeOwn_L ==> CAT_H	12	41
287	SingleParentF_H Male17-24_H ==> CAT_H	6	41
288	HeterogeneityInx_H Employment_H ==> CAT_MH	6	41

APPENDIX D: (continued)

289	SingleParentF_H AfricanA_H HEdu_M Income_L ==> CAT_H	5	41
290	HeterogeneityInx_H HEdu_M Income_L ==> CAT_MH	6	41
291	UnstableRent_H HEdu_M Income_L ==> CAT_MH	5	41
292	UnstableRent_H ==> CAT_H	11	40
293	Male17-24_H HEdu_M Income_L ==> CAT_H	5	40
294	SingleParentF_L AfricanA_L MUHouse_M ==> CAT_MH	6	40

APPENDIX E: REPRESENTATIVE MINED FUZZY SARS FOR DANGEROUS
STREET DUE TO MOTOR VEHICLE THEFT USING SUPPORT THRESHOLD OF
5%, CONFIDENCE THRESHOLD OF 40%

Rules	S%	C%
HeterogeneityInx_H UnstableRent_H Income_L ==> MVT_H	5	65
HeterogeneityInx_H UnstableRent_H MUHouse_H ==> MVT_H	5	63
HeterogeneityInx_H UnstableRent_H ==> MVT_H	6	60
HeterogeneityInx_H MUHouse_H ==> MVT_H	6	56
Male17-24_H MUHouse_H ==> MVT_H	6	51
Male17-24_H UnstableRent_H Income_L ==> MVT_H	5	51
UnstableRent_H HEdu_M MUHouse_H ==> MVT_H	5	51
UnstableRent_H MUHouse_H Income_L ==> MVT_H	6	48
Male17-24_H UnstableRent_H ==> MVT_H	6	47
Male17-24_H HeterogeneityInx_H Income_L ==> MVT_H	6	46
HEdu_M MUHouse_H ==> MVT_H	6	45
Male17-24_H HeterogeneityInx_H ==> MVT_H	6	44
MUHouse_H Income_L ==> MVT_H	6	43
SingleParentF_H UnstableRent_H ==> MVT_H	5	43
SingleParentF_H UnstableRent_H Income_L ==> MVT_H	5	43
HomeOwn_L UnstableRent_H MUHouse_H ==> MVT_H	6	42
UnstableRent_H HEdu_M Income_L ==> MVT_H	5	42
HomeOwn_L MUHouse_H ==> MVT_H	6	40

APPENDIX F: REPRESENTATIVE MINED FUZZY SARS FOR DANGEROUS
STREET DUE TO THEFT FROM MOTOR VEHICLE USING SUPPORT
THRESHOLD OF 5%, CONFIDENCE THRESHOLD OF 40%

	Rules	S%	C%
1	HomeOwn_MH AfricanA_H ==> TFM_MH	8	74
2	AfricanA_H MUHouse_M ==> TFM_MH	7	74
3	SingleParentF_MH UnstableRent_M ==> TFM_MH	8	74
4	SingleParentF_MH HEdu_M ==> TFM_MH	10	72
5	UnstableRent_L Income_L ==> TFM_MH	10	72
6	HeterogeneityInx_L Income_L ==> TFM_MH	6	72
7	SingleParentF_MH Male17-24_L ==> TFM_MH	6	72
8	SingleParentF_MH HeterogeneityInx_H ==> TFM_MH	6	71
9	AfricanA_H Employment_L ==> TFM_MH	6	71
10	SingleParentF_MH MUHouse_M ==> TFM_MH	6	71
11	HeterogeneityInx_L UnstableRent_M ==> TFM_MH	7	71
12	HomeOwn_MH HeterogeneityInx_H ==> TFM_MH	8	71
13	Employment_L HEdu_M ==> TFM_MH	6	70
14	AfricanA_H UnstableRent_M ==> TFM_MH	6	70
15	AfricanA_M MUHouse_L ==> TFM_MH	5	70
16	Male17-24_MH UnstableRent_M ==> TFM_MH	8	70
17	Employment_L Income_L ==> TFM_MH	9	70
18	SingleParentF_H Employment_L ==> TFM_MH	6	70
19	MUHouse_L Income_L ==> TFM_MH	14	70
20	HomeOwn_MH MUHouse_M ==> TFM_MH	12	69
21	HomeOwn_MH UnstableRent_M ==> TFM_MH	13	69
22	SingleParentF_MH Employment_H ==> TFM_MH	8	69
23	Male17-24_MH HEdu_M ==> TFM_MH	9	69
24	AfricanA_M HEdu_M ==> TFM_MH	7	69
25	HomeOwn_MH MUHouse_L ==> TFM_MH	7	69
26	HEdu_L Income_H ==> TFM_MH	6	68
27	HomeOwn_MH HEdu_M ==> TFM_MH	14	68
28	MUHouse_M Income_L ==> TFM_MH	11	68
29	HomeOwn_MH Income_L ==> TFM_MH	15	68
30	HeterogeneityInx_H AfricanA_M ==> TFM_MH	6	68
31	SingleParentF_MH HomeOwn_MH ==> TFM_MH	8	68
32	HomeOwn_L Employment_L ==> TFM_MH	5	68
33	Male17-24_MH AfricanA_H ==> TFM_MH	6	68
34	Male17-24_L Income_L ==> TFM_MH	9	68
35	Male17-24_MH MUHouse_L ==> TFM_MH	7	67
36	Employment_L ==> TFM_MH	10	67

APPENDIX F: (continued)

37	SingleParentF_L Income_L ==> TFM_MH	5	67
38	SingleParentF_MH MUHouse_L ==> TFM_MH	7	67
39	HomeOwn_MH Male17-24_L ==> TFM_MH	11	67
40	SingleParentF_H HomeOwn_MH ==> TFM_MH	7	67
41	SingleParentF_MH Income_L ==> TFM_MH	12	67
42	UnstableRent_M Income_L ==> TFM_MH	12	66
43	HomeOwn_MH UnstableRent_L ==> TFM_MH	7	66
44	SingleParentF_MH ==> TFM_MH	16	66
45	Male17-24_MH MUHouse_M ==> TFM_MH	6	66
46	HomeOwn_MH HeterogeneityInx_L ==> TFM_MH	7	66
47	HeterogeneityInx_L MUHouse_M ==> TFM_MH	8	66
48	HomeOwn_MH ==> TFM_MH	26	66
49	Employment_H MUHouse_M ==> TFM_MH	9	65
50	HomeOwn_MH Employment_M ==> TFM_MH	10	65
51	SingleParentF_H MUHouse_M ==> TFM_MH	5	65
52	HomeOwn_MH Male17-24_MH ==> TFM_MH	8	65
53	UnstableRent_M MUHouse_M ==> TFM_MH	12	65
54	UnstableRent_M HEdu_M ==> TFM_MH	11	65
55	UnstableRent_M ==> TFM_MH	21	65
56	SingleParentF_H MUHouse_L ==> TFM_MH	7	65
57	HomeOwn_MH Male17-24_H ==> TFM_MH	6	65
58	Employment_H UnstableRent_M ==> TFM_MH	10	65
59	Employment_H Income_L ==> TFM_MH	11	65
60	HEdu_M MUHouse_L ==> TFM_MH	16	65
61	SingleParentF_MH Employment_M ==> TFM_MH	5	65
62	HomeOwn_H Income_L ==> TFM_MH	6	64
63	HomeOwn_MH Employment_H ==> TFM_MH	12	64
64	HeterogeneityInx_H UnstableRent_L ==> TFM_MH	8	64
65	HeterogeneityInx_H MUHouse_L ==> TFM_MH	10	64
66	UnstableRent_M MUHouse_L ==> TFM_MH	5	64
67	HeterogeneityInx_MH HEdu_M ==> TFM_MH	13	64
68	SingleParentF_H MUHouse_H ==> TFM_H	5	64
69	SingleParentF_MH AfricanA_M ==> TFM_MH	7	64
70	HeterogeneityInx_H Employment_H ==> TFM_MH	9	64
71	Male17-24_L HeterogeneityInx_MH ==> TFM_MH	8	63
72	HeterogeneityInx_MH MUHouse_L ==> TFM_MH	8	63
73	MUHouse_M ==> TFM_MH	20	63
74	AfricanA_M ==> TFM_MH	13	63
75	SingleParentF_L HomeOwn_MH ==> TFM_MH	10	63
76	HomeOwn_MH AfricanA_L ==> TFM_MH	12	63
77	AfricanA_H Income_L ==> TFM_MH	16	63

APPENDIX F: (continued)

78	Employment_H UnstableRent_L ==> TFM_MH	12	63
79	HeterogeneityInx_MH UnstableRent_M ==> TFM_MH	7	63
80	HeterogeneityInx_MH Income_L ==> TFM_MH	12	63
81	Male17-24_MH Employment_H ==> TFM_MH	6	63
82	AfricanA_H MUHouse_L ==> TFM_MH	8	63
83	Male17-24_MH Income_L ==> TFM_MH	11	63
84	AfricanA_H UnstableRent_L ==> TFM_MH	5	63
85	Employment_H MUHouse_L ==> TFM_MH	11	62
86	Male17-24_MH Employment_M ==> TFM_MH	6	62
87	UnstableRent_M HEdu_L ==> TFM_MH	7	62
88	HomeOwn_MH HeterogeneityInx_MH ==> TFM_MH	11	62
89	Male17-24_MH ==> TFM_MH	16	62
90	HEdu_L MUHouse_M ==> TFM_MH	6	62
91	Male17-24_MH HeterogeneityInx_MH ==> TFM_MH	8	62
92	SingleParentF_MH Male17-24_MH ==> TFM_MH	6	62
93	HomeOwn_MH HEdu_L ==> TFM_MH	10	62
94	Male17-24_L Income_H ==> TFM_MH	8	62
95	HomeOwn_MH AfricanA_M ==> TFM_MH	6	62
96	HeterogeneityInx_H UnstableRent_M ==> TFM_MH	6	62
97	AfricanA_M Income_L ==> TFM_MH	9	62
98	HEdu_M Income_L ==> TFM_MH	20	62
99	HomeOwn_H Employment_H ==> TFM_MH	11	61
100	Male17-24_L MUHouse_M ==> TFM_MH	9	61
101	HeterogeneityInx_L HEdu_L ==> TFM_MH	10	61
102	HEdu_M MUHouse_M ==> TFM_MH	12	61
103	Male17-24_L Employment_H ==> TFM_MH	14	61
104	Employment_M UnstableRent_M ==> TFM_MH	8	61
105	HomeOwn_MH Income_M ==> TFM_MH	7	61
106	Income_L ==> TFM_MH	31	61
107	SingleParentF_L Income_H ==> TFM_MH	9	61
108	HomeOwn_H HeterogeneityInx_H ==> TFM_MH	8	61
109	Male17-24_L UnstableRent_M ==> TFM_MH	9	61
110	Income_H ==> TFM_MH	9	61
111	SingleParentF_MH HeterogeneityInx_MH ==> TFM_MH	8	61
112	AfricanA_L UnstableRent_M ==> TFM_MH	11	61
113	AfricanA_L Income_H ==> TFM_MH	9	61
114	AfricanA_L Employment_H ==> TFM_MH	17	61
115	UnstableRent_L HEdu_M ==> TFM_MH	14	61
116	HeterogeneityInx_MH MUHouse_M ==> TFM_MH	8	60
117	AfricanA_H ==> TFM_MH	17	60
118	AfricanA_H HEdu_M ==> TFM_MH	12	60

APPENDIX F: (continued)

119	SingleParentF_L UnstableRent_M ==> TFM_MH	9	60
120	Male17-24_L HEdu_M ==> TFM_MH	17	60
121	Employment_H ==> TFM_MH	27	60
122	SingleParentF_L Employment_H ==> TFM_MH	16	60
123	MUHouse_L ==> TFM_MH	25	60
124	Employment_H HEdu_M ==> TFM_MH	15	60
125	Employment_H HEdu_L ==> TFM_MH	10	60
126	HomeOwn_L HeterogeneityInx_H ==> TFM_H	5	59
127	HeterogeneityInx_MH UnstableRent_L ==> TFM_MH	6	59
128	HEdu_L Income_L ==> TFM_MH	9	59
129	HeterogeneityInx_MH ==> TFM_MH	20	59
130	HeterogeneityInx_MH AfricanA_M ==> TFM_MH	6	59
131	HomeOwn_H MUHouse_L ==> TFM_MH	14	59
132	HeterogeneityInx_L Income_H ==> TFM_MH	8	59
133	HomeOwn_MH UnstableRent_H ==> TFM_MH	6	58
134	SingleParentF_L MUHouse_M ==> TFM_MH	9	58
135	SingleParentF_H AfricanA_H ==> TFM_MH	11	58
136	HEdu_M ==> TFM_MH	33	58
137	Male17-24_MH AfricanA_L ==> TFM_MH	5	58
138	SingleParentF_L HeterogeneityInx_H ==> TFM_MH	5	58
139	HeterogeneityInx_MH AfricanA_L ==> TFM_MH	10	58
140	HeterogeneityInx_L ==> TFM_MH	20	58
141	AfricanA_M Employment_H ==> TFM_MH	6	58
142	HeterogeneityInx_L Employment_H ==> TFM_MH	8	58
143	UnstableRent_L MUHouse_L ==> TFM_MH	19	58
144	HeterogeneityInx_MH Employment_H ==> TFM_MH	9	58
145	SingleParentF_L HeterogeneityInx_MH ==> TFM_MH	8	58
146	HEdu_L MUHouse_H ==> TFM_MH	6	58
147	UnstableRent_M Income_M ==> TFM_MH	6	58
148	Male17-24_L ==> TFM_MH	28	58
149	HomeOwn_MH MUHouse_H ==> TFM_MH	6	58
150	SingleParentF_L HEdu_L ==> TFM_MH	12	58
151	HomeOwn_H UnstableRent_L ==> TFM_MH	15	58
152	UnstableRent_L ==> TFM_MH	23	58
153	AfricanA_L MUHouse_M ==> TFM_MH	10	57
154	Employment_M MUHouse_M ==> TFM_MH	7	57
155	AfricanA_L Income_L ==> TFM_MH	7	57
156	HomeOwn_L HeterogeneityInx_L ==> TFM_MH	6	57
157	HomeOwn_H HEdu_M ==> TFM_MH	12	57
158	MUHouse_M Income_M ==> TFM_MH	6	57
159	Male17-24_L UnstableRent_L ==> TFM_MH	15	56

APPENDIX F: (continued)

160	HomeOwn_L HeterogeneityInx_MH ==> TFM_MH	6	56
161	HEdu_L ==> TFM_MH	20	56
162	HomeOwn_H ==> TFM_MH	18	56
163	AfricanA_L HEdu_L ==> TFM_MH	12	56
164	Male17-24_L HeterogeneityInx_H ==> TFM_MH	6	56
165	Male17-24_L MUHouse_L ==> TFM_MH	14	56
166	HeterogeneityInx_H ==> TFM_MH	18	56
167	Male17-24_H Income_L ==> TFM_MH	12	56
168	HeterogeneityInx_MH Employment_M ==> TFM_MH	7	56
169	HeterogeneityInx_H Income_L ==> TFM_MH	13	56
170	Male17-24_L HeterogeneityInx_L ==> TFM_MH	14	56
171	HeterogeneityInx_L HEdu_M ==> TFM_MH	9	56
172	SingleParentF_L ==> TFM_MH	27	56
173	Male17-24_L HEdu_L ==> TFM_MH	11	56
174	SingleParentF_H Income_L ==> TFM_MH	14	55
175	SingleParentF_L MUHouse_H ==> TFM_MH	6	55
176	SingleParentF_L HomeOwn_H ==> TFM_MH	12	55
177	HeterogeneityInx_H MUHouse_H ==> TFM_H	6	55
178	Male17-24_L AfricanA_L ==> TFM_MH	20	55
179	HeterogeneityInx_MH UnstableRent_H ==> TFM_MH	6	55
180	Employment_M MUHouse_L ==> TFM_MH	9	55
181	HeterogeneityInx_H HEdu_M ==> TFM_MH	11	55
182	HomeOwn_H Male17-24_L ==> TFM_MH	13	54
183	Employment_H Income_M ==> TFM_MH	12	54
184	Male17-24_H Employment_H ==> TFM_MH	6	54
185	AfricanA_L MUHouse_L ==> TFM_MH	12	54
186	SingleParentF_L AfricanA_L ==> TFM_MH	23	54
187	SingleParentF_L UnstableRent_L ==> TFM_MH	14	54
188	SingleParentF_L Male17-24_L ==> TFM_MH	19	54
189	AfricanA_L ==> TFM_MH	28	54
190	Male17-24_H MUHouse_H ==> TFM_H	6	54
191	SingleParentF_L HEdu_M ==> TFM_MH	14	54
192	SingleParentF_H ==> TFM_MH	14	54
193	HeterogeneityInx_H AfricanA_H ==> TFM_MH	8	54
194	HEdu_M Income_M ==> TFM_MH	10	54
195	SingleParentF_L MUHouse_L ==> TFM_MH	12	54
196	HeterogeneityInx_L AfricanA_L ==> TFM_MH	14	54
197	HeterogeneityInx_H UnstableRent_H ==> TFM_H	5	54
198	SingleParentF_L HeterogeneityInx_L ==> TFM_MH	14	54
199	HomeOwn_H AfricanA_L ==> TFM_MH	12	53
200	AfricanA_H Employment_M ==> TFM_MH	6	53

APPENDIX F: (continued)

201	SingleParentF_H HEdu_M ==> TFM_MH	9	53
202	Employment_M Income_L ==> TFM_MH	12	53
203	Male17-24_H AfricanA_H ==> TFM_MH	7	53
204	Male17-24_H ==> TFM_MH	14	53
205	HomeOwn_H HEdu_L ==> TFM_MH	5	53
206	AfricanA_L UnstableRent_L ==> TFM_MH	14	53
207	HomeOwn_L MUHouse_H ==> TFM_H	8	53
208	AfricanA_L HEdu_M ==> TFM_MH	14	53
209	Employment_M HEdu_M ==> TFM_MH	13	53
210	Male17-24_H HeterogeneityInx_H ==> TFM_MH	8	53
211	HeterogeneityInx_MH Income_M ==> TFM_MH	7	52
212	UnstableRent_L HEdu_L ==> TFM_MH	8	52
213	HomeOwn_H HeterogeneityInx_L ==> TFM_MH	7	52
214	HeterogeneityInx_L Employment_M ==> TFM_MH	7	52
215	Employment_M ==> TFM_MH	21	52
216	AfricanA_H UnstableRent_H ==> TFM_MH	6	52
217	HeterogeneityInx_L UnstableRent_L ==> TFM_MH	9	52
218	HomeOwn_L Income_L ==> TFM_MH	10	51
219	HEdu_L MUHouse_L ==> TFM_MH	8	51
220	Income_M ==> TFM_MH	17	51
221	HomeOwn_L AfricanA_H ==> TFM_MH	5	51
222	HeterogeneityInx_L MUHouse_L ==> TFM_MH	7	51
223	Employment_H MUHouse_H ==> TFM_MH	7	51
224	HomeOwn_H Income_M ==> TFM_MH	8	51
225	HomeOwn_L Male17-24_H ==> TFM_H	5	50
226	Male17-24_L Employment_M ==> TFM_MH	10	50
227	Employment_M HEdu_L ==> TFM_MH	6	50
228	Employment_M UnstableRent_L ==> TFM_MH	8	50
229	HeterogeneityInx_MH HEdu_L ==> TFM_MH	5	50
230	UnstableRent_H ==> TFM_MH	14	50
231	AfricanA_L MUHouse_H ==> TFM_MH	6	49
232	Male17-24_L Income_M ==> TFM_MH	11	49
233	SingleParentF_H Employment_M ==> TFM_MH	6	49
234	UnstableRent_H Income_L ==> TFM_MH	10	49
235	HomeOwn_H Employment_M ==> TFM_MH	6	49
236	SingleParentF_L Income_M ==> TFM_MH	13	49
237	MUHouse_L Income_M ==> TFM_MH	8	49
238	AfricanA_L Income_M ==> TFM_MH	12	49
239	Employment_H UnstableRent_H ==> TFM_MH	6	49
240	UnstableRent_L Income_M ==> TFM_MH	8	49
241	HomeOwn_L ==> TFM_MH	14	48

APPENDIX F: (continued)

242	HeterogeneityInx_H Employment_M ==> TFM_MH	7	48
243	SingleParentF_L Employment_M ==> TFM_MH	9	48
244	HEdu_M MUHouse_H ==> TFM_H	6	48
245	MUHouse_H ==> TFM_MH	12	48
246	UnstableRent_H HEdu_M ==> TFM_MH	7	47
247	SingleParentF_H HeterogeneityInx_H ==> TFM_MH	7	47
248	UnstableRent_H MUHouse_H ==> TFM_H	9	47
249	Male17-24_H HEdu_M ==> TFM_MH	7	47
250	HeterogeneityInx_L Income_M ==> TFM_MH	5	47
251	SingleParentF_L HomeOwn_L ==> TFM_MH	5	46
252	HomeOwn_L HEdu_M ==> TFM_MH	7	46
253	Employment_M Income_M ==> TFM_MH	5	46
254	HomeOwn_L UnstableRent_H ==> TFM_MH	8	46
255	SingleParentF_H Male17-24_H ==> TFM_MH	6	46
256	Male17-24_H UnstableRent_H ==> TFM_MH	6	45
257	AfricanA_L Employment_M ==> TFM_MH	10	45
258	Male17-24_H UnstableRent_H ==> TFM_H	6	44
259	MUHouse_H Income_L ==> TFM_H	7	44
260	HomeOwn_L UnstableRent_H ==> TFM_H	8	44
261	SingleParentF_H UnstableRent_H ==> TFM_MH	5	44
262	HEdu_L Income_M ==> TFM_MH	5	44
263	SingleParentF_H UnstableRent_H ==> TFM_H	5	44
264	HEdu_M MUHouse_H ==> TFM_MH	5	42
265	UnstableRent_H MUHouse_H ==> TFM_MH	8	41
266	MUHouse_H Income_L ==> TFM_MH	6	41
267	Employment_H MUHouse_H ==> TFM_H	5	40
